DISEASE VECTORS OF PAPUA NEW GUINEA, MEMBERS OF THE *ANOPHELES PUNCTULATUS* SPECIES COMPLEX (DIPTERA: CULICIDAE)—MOLECULAR DIVERSITY, SPECIES IDENTIFICATION AND IMPLICATIONS FOR INTEGRATED VECTOR MANAGEMENT

by

CARA NICOLE HALLDIN

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy

Dissertation Advisor: Dr. Peter A. Zimmerman, Ph.D.

Department of Biology
CASE WESTERN RESERVE UNIVERSITY

May, 2011
We hereby approve the dissertation of

Cara Nicole Halldin

candidate for the Doctor of Philosophy degree*.

(signed) Jean H. Burns, PhD

(chair of the committee)

Joseph B. Keiper, PhD

Joseph F. Koonce, PhD

Michael F. Benard, PhD

Charles H. King, MD

Peter A. Zimmerman, PhD

(date) March 4th, 2011

*We also certify that written approval has been obtained for any proprietary material contained therein.
# Table of Contents

### Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xi</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>xiii</td>
</tr>
<tr>
<td>Abstract</td>
<td>1</td>
</tr>
</tbody>
</table>

## I. Chapter 1. *Anopheles punctulatus* species complex: The Status of Disease Vectors in Papua New Guinea

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Introduction: Malaria Transmission and the <em>Anopheles</em> vector</td>
<td>4</td>
</tr>
<tr>
<td>B. Australo-Papuan Region</td>
<td>7</td>
</tr>
<tr>
<td>1. Geological and Zoological History</td>
<td>7</td>
</tr>
<tr>
<td>2. Geography and Climate</td>
<td>11</td>
</tr>
<tr>
<td>C. <em>Anopheles</em> of Papua New Guinea</td>
<td>13</td>
</tr>
<tr>
<td>1. Historical Entomological Accounts</td>
<td>13</td>
</tr>
<tr>
<td>2. Species discrimination by molecular taxonomy</td>
<td>17</td>
</tr>
<tr>
<td>3. Species distribution and larval habitat preference</td>
<td>20</td>
</tr>
<tr>
<td>a. <em>An. farauti</em> s.s.</td>
<td>22</td>
</tr>
<tr>
<td>b. <em>An. hinesorum</em></td>
<td>22</td>
</tr>
<tr>
<td>c. <em>An. farauti</em> 4</td>
<td>23</td>
</tr>
<tr>
<td>d. <em>An. punctulatus</em></td>
<td>23</td>
</tr>
<tr>
<td>e. <em>An. koliensis</em></td>
<td>24</td>
</tr>
<tr>
<td>f. Location of minor <em>An. punctulatus</em> complex members found in PNG</td>
<td>24</td>
</tr>
<tr>
<td>4. Understanding relationships of AP sibling species</td>
<td>25</td>
</tr>
<tr>
<td>D. <em>Anopheles</em> and <em>Plasmodium</em></td>
<td>29</td>
</tr>
<tr>
<td>1. Human migrations and introduction of malaria into New Guinea</td>
<td>29</td>
</tr>
<tr>
<td>2. Host feeding preferences</td>
<td>31</td>
</tr>
<tr>
<td>3. Vectorial capacity</td>
<td>33</td>
</tr>
<tr>
<td>4. Early Malaria Control in PNG</td>
<td>35</td>
</tr>
<tr>
<td>E. Depth of <em>An. punctulatus</em> complex published work</td>
<td>37</td>
</tr>
<tr>
<td>F. Dissertation Aims</td>
<td>39</td>
</tr>
</tbody>
</table>

## II. Chapter 2. Rationale for Study Design and Study Methods

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Introduction</td>
<td>44</td>
</tr>
<tr>
<td>B. Mosquito Collections, Morphological Identification and Genomic DNA extraction</td>
<td>45</td>
</tr>
<tr>
<td>1. Mosquito collections</td>
<td>45</td>
</tr>
<tr>
<td>2. Morphological species identification</td>
<td>47</td>
</tr>
</tbody>
</table>
III. Chapter 3. Genetic Analysis of ITS2 Sequence: Evaluating Species Differences 68

A. Introduction: *Anopheles punctulatus* Sibling Species Definitions 69
B. Methods 71
C. Results 72
   1. ITS2 Sequence Analysis of AP Sibling Species 72
   2. Development of High-throughput Species Diagnostic Method Utilizing the ITS2 Sequence 79
   3. Application of ITS2 LDR-FMA 83
      a. Evaluation of species identification with established method 83
      b. Multiplex evaluation of mosquito species and identification with *W. bancrofti* detection 85
D. Discussion 86

IV. Chapter 4. Interrogation of the Voltage Gated Sodium Channel Gene: Evaluating Species Differences and Monitoring Mutations Associated with Insecticide Resistance 92

A. Introduction: *VGSC*, Insecticide Resistance Mutations and the *An. punctulatus* Complex 93
B. Methods 98
C. Results 98
   1. Sequence Analysis of *VGSC* for *An. punctulatus* Sibling Species 98
      a. General Sequence Observations 100
      b. Coding region sequence comparisons 101
         i. General comparisons 101
         ii. SNP identification 103
      c. Observations of the Introns 107
   2. Evaluation of Intronic Patterns and Simultaneous Monitoring for kdr Mutations 110
      a. *VGSC* LDR-FMA multiplex design 111
   3. Application of the *VGSC* multiplex assay 115
      a. *VGSC* LDR-FMA analysis of sample set previously subjected to insecticide susceptibility phenotype evaluation 116
D. Discussion

V. Chapter 5. Evaluating Genetic Diversity of the *Anopheles punctulatus* species Complex by Investigating Mitochondrial Genes 123
   A. Introduction: Mitochondrial DNA- Advantages, disadvantages, and what is known about the AP complex mitochondrial sequence 124
   B. Methods 129
   C. Results
      1. Sequencing of mtDNA for *An. punctulatus* Sibling Species 130
         a. General mtDNA sequence observations 132
         b. *COI* diversity within the AP complex 137
         c. Closer examination of *COI* sequences 138
         d. Diversity of *Cytb* sequence as a complimentary mtDNA sequence 141
         e. Differentiation of the AP complex: How does the differentiation of the AP complex compare with other cryptic *Anopheles* groups? 143
         f. Intra-species mismatch distribution predictive of phylogenetic signal 144
   D. Discussion 148

VI. Chapter 6. Using Phylogenetic Analysis to Assess Relationships Among *An. punctulatus* species complex Members 153
   A. Introduction: A Phylogenetic Approach to Understanding Inter-*An. punctulatus* Group Relationships 154
   B. Methods 158
   C. Results
      1. Neighbor-Joining analysis 163
         a. Nuclear sequence results 164
         b. Mitochondrial sequence results 166
         c. Phylogeographical observations within of *An. farauti* s.s. and *An. hinesorum* 169
         d. Summary of Neighbor-Joining results 171
      2. Maximum Likelihood Analysis 173
         a. Nuclear sequence results 173
         b. Mitochondrial sequence results 173
      3. Gene Trees to Species Tree: RAxML simultaneous analysis of three genes to infer species tree 175
      4. Approaches to estimate divergence times 177
   D. Discussion 180
      1. Phylogeographic patterns 180
      2. Phylogenetic relationships within the AP sibling species 182
      3. Observations of ‘rogue’ individuals 184
      4. Phenotypes and phylogenetic relationships 185
      5. Divergence time of AP sibling species 186

VII. Chapter 7. Project Summary 189
   A. Summary of Project Findings 190
1. Genetic diversity supporting multiple species hypothesis and defining a baseline understanding of AP species diversity and differentiation 192
2. Molecular tools 196

B. Integrated Vector Management: Contributions of this Project 198
1. IVM in PNG 199
2. IVM for other Anopheles species complexes 202

C. Malaria Elimination and Importance of Molecular Epidemiology 203

D. Future Directions of Research Associated with this Project 205

VIII. Appendix A. Sequence Alignments of Four Genes Used in this Dissertation 209
IX. References 214
List of Tables

Chapter 1
Table 1.1 Literature Search of Malaria Vectors 38

Chapter 2
Table 2.1 Mosquito Collection Site Data 46
Table 2.2 PCR Primers and Amplification Conditions 55
Table 2.3 Ligation detection reaction probes for PNG Anopheles species identification and mutation detection 56

Chapter 3
Table 3.1 Mosquito Collections for ITS2 Evaluation 72
Table 3.2 ITS2 Sequencing Results 74
Table 3.3 Concordance assessment between morphology and ITS2 classification for An. punctulatus sibling species 84
Table 3.4 Concordance assessment of molecular ITS2 classification methods: RFLP vs. LDR-FMA for An. punctulatus sibling species 84
Table 3.5 Dreikikir Area Mosquito Survey 85

Chapter 4
Table 4.1 Mosquito Collections for VGSC Evaluation 99
Table 4.2 VGSC Sequencing Results 100
Table 4.3 An. punctulatus species complex VGSC coding sequence comparisons 101
Table 4.4 VGSC Exon Single Nucleotide Polymorphisms found in An. punctulatus species complex members 104
Table 4.5 An. punctulatus species complex VGSC intron sequence comparisons 108
Table 4.6 Concordance assessment of molecular ITS2 and VGSC LDR-FMA classification methods for An. punctulatus species complex members 116

Chapter 5
Table 5.1 Mosquito Collections for COI and Cytb Sequencing 130
Table 5.2 Pairwise Percent Identity of mtDNA sequences among Anopheles species 132
Table 5.3 COI Sequencing Results 136
Table 5.4 Cytb Sequencing Results 136
Table 5.5 Summary Statistics for *An. punctulatus* species group COI Diversity from pruned alignment

Table 5.6 Summary Statistics for *An. punctulatus* species group *Cytb* Diversity from pruned alignment

Table 5.7 Interspecific Pairwise Estimates for $F_{st}$ for COI Sequences

Chapter 6

Table 6.1 Details of DNA Sequence Alignments

Table 6.2 Genbank Accession Information
List of Figures

Chapter 1

Figure 1.1 *Plasmodium* lifecycle stages in the human host and *Anopheles* vector 4
Figure 1.2 Phylogenetic relationships of the Culicidae Family 6
Figure 1.3 Papua New Guinea Provincial boundaries and key locations 8
Figure 1.4 Figure adapted from Mayr (1944) demonstrating location of Wallace, Weber and Lydekker Lines 10
Figure 1.5 Papua New Guinea general maximum elevation and main geographic features 12
Figure 1.6 Morphological characteristics used for species differentiation among *An. punctulatus*, *An. farauti* s.l. and *An. koliensis* 14
Figure 1.7 ITS2 PCR-RFLP patterns observed for 10 *An. punctulatus* complex members 19
Figure 1.8 Collection locations from Cooper et al. 2002 survey of 795 throughout PNG 20
Figure 1.9 General distribution of the five most common AP complex species 21
Figure 1.10 General representation of current phylogenetic hypotheses adapted from Beebe (2000) 27

Chapter 2

Figure 2.1 Locations of mosquito collections in Papua New Guinea 45
Figure 2.2 Individual mosquitoes collected and analyzed for this dissertation 46
Figure 2.3 Wing and proboscis morphological key for female *An. farauti* s.l., *An. punctulatus*, and *An. koliensis*. 47
Figure 2.4 Basic design of a ligation detection reaction-fluorescent microsphere assay (LDR-FMA) 58
Figure 2.5 Example demonstrating two different models of substitution 62

Chapter 3

Figure 3.1 ITS2 reference sequences compared with consensus sequences from the current project 76
Figure 3.2 Alignment of *An. farauti* s.s. representatives 78
Figure 3.3 ITS2 consensus alignment demonstrating locations of LDR-FMA species identification probes 81
Figure 3.4 Detection of species-specific DNAs by ITS2 LDR-FMA

Chapter 4
Figure 4.1 Arthropod Voltage Gated Sodium Channel and region of interest
Figure 4.2 Amplified region of VGSC with regions of interest
Figure 4.3 Detection of species-specific DNAs and kdr mutations by VGSC LDR-FMA

Chapter 5
Figure 5.1 Example of observed outliers in mtDNA sequence analysis
Figure 5.2 Screenshot of whole sequence alignments of An. hinesorum Cytb demonstrating sequence differences
Figure 5.3 Alignment of COI portion demonstrating outliers are more similar to other species within the AP complex
Figure 5.4 Mismatch distributions

Chapter 6
Figure 6.1 Neighbor-joining phylogenetic trees
Figure 6.2 Schematic demonstrating number of individual mosquitoes sequenced for multiple genes
Figure 6.3 Collection locations of An. farauti s.s. and An. hinesorum described
Figure 6.4 Neighbor-joining results for ITS2 alignment adding remaining AP species
Figure 6.5 Maximum likelihood phylogenetic trees
Figure 6.6 Maximum likelihood analysis using RAxML and COI
Figure 6.7 BEAST chronogram using COI

Appendix A
Figure A.1 ITS2 Alignment
Figure A.2 VGSC Alignment
Figure A.3 COI Alignment
Figure A.4 Cytb Alignment
Acknowledgments

In reflecting on the last four years, many people have contributed countless hours to ensure my achievement both at the bench and in the classroom. I would like to express my gratitude first and foremost to my graduate advisor and mentor, Peter A. Zimmerman, Ph.D. for his patience, guidance and instruction throughout my graduate career. Pete’s constant encouragement to dig deeper and keep an open mind has helped me mature both as a scientist and as a person.

It is also a pleasure for me to thank members of my committee, Mike Benard, Ph.D., Joe Keiper, Ph.D., Charlie King, MD, and Joe Koonce, Ph.D.—who have provided intellectual support and helpful criticism throughout my graduate career. I consider myself lucky to have assembled a group who exhibit incredible enthusiasm towards my research and have provided sincere encouragement as I progressed through this dissertation project.

Many members of the Zimmerman Laboratory, both past and present, have provided training, assistance, comic relief and moral support over the past couple years. I must thank Rajeev Mehlotra, Ph.D. for his comments on my manuscripts and parts of this dissertation, but most importantly for fielding my endless questions and providing helpful instruction, sound advice, and constructive criticism. I also thank Brian Grimberg, Ph.D. and Jeana DaRe Ph.D. for always contributing their advice with a side of wit and perhaps a dash of harsh sarcasm, a true sign of friendship. This research was made possible by Lisa Reimer, Ph.D., Eddie Thomsen and John Bosco Keven who have provided field collections, data and practical
instruction both during my time in Cleveland and my stay in Madang. The laboratory research would not have been possible without the instruction and tireless assistance of Dave McNamara, Melinda Zikursh, and Laurie Gray. And I cannot forget the undergraduates—Kogulan Nadesakumaran and Allison Zimmerman have worked diligently towards generating data for this project and have always provided stimulating lab conversation.

Finally, I would not be here without my family's support, encouragement and love. No one has supported me more over the years than my mother. She has exemplified what it means to be a strong, successful person. Persevering through the challenges life has presented her, she has taught me to think positively and to never give up. And, last but certainly not least, Jesse, my husband, who, God love him, has ‘endured' Cleveland and has lovingly provided motivation, support and endless laughter (oh, and good health insurance too).

I think a quote from Mark Twain summarizes my relief: "I'm glad I did it, partly because it was worth it, but mostly because I shall never have to do it again."—my sincere gratitude to all that have helped me to accomplish this milestone in my life.
**List of Abbreviations**

- **AFs.s. or AF1**: *Anopheles farauti* sensu stricto (formerly *An. farauti* 1)
- **AF4**: *An. farauti* 4
- **AH or AF2**: *An. hinesorum* (formerly *An. farauti* 2)
- **AK**: *An. koliensis*
- **AP**: *An. punctulatus*
- **BOLD**: Barcode of Life Datasystems
- **CBOL**: Consortium for the Barcode of Life
- **COI**: Cytochrome Oxidase I gene
- **Cytb**: Cytochrome B gene
- **DDT**: dichlorodiphenyltrichloroethane insecticide
- **GPS**: global positioning system
- **IRS**: indoor residual spraying
- **ITN**: insecticide treated bednet
- **ITS2**: Internal Transcribed Spacer 2 (rDNA)
- **IVM**: Integrated Vector Management
- **LDR-FMA**: ligation detection reaction-fluorescent microsphere assay
- **LLIN**: long-lasting insecticidal net
- **kdr**: knock down resistance mutation
- **malERA**: Malaria Eradication Research Agenda
- **MYA**: million years ago
- **NGO**: Non-governmental Organizations
- **PCR**: polymerase chain reaction
- **PNG**: Papua New Guinea
- **PNGIMR**: Papua New Guinea Institute of Medical Research
- **RFLP**: restriction fragment length polymorphism
- **SNP**: single nucleotide polymorphism
- **STT**: saline tolerance test
- **VGSC**: Voltage Gated Sodium Channel gene
- **WHO**: World Health Organization
Disease vectors of Papua New Guinea, members of the *Anopheles punctulatus* Species Complex (Diptera: Culicidae)—Molecular diversity, species identification and implications for integrated vector management

Abstract

by

CARA NICOLE HALLDIN

Members of the *Anopheles punctulatus* species complex are thought to be the predominant vectors of malaria and filariasis in Papua New Guinea (PNG). Establishing the importance of each species as a disease vector requires a thorough understanding of each species’ general ecology, biting habits, larval habitat preferences, insecticide susceptibility and capacity to obtain/propagate parasitic infections. But first, reliable species definitions and identification methods must be established. Prior this dissertation little understanding of the genetic diversity within and among these species was available. Here, by sequencing two nuclear and two mitochondrial genes I have examined species definitions and relationships for the five most common species within the group (*An. punctulatus, An. koliensis, An. farauti s.s., An. hinesorum* and *An. farauti 4*) to better characterize species diversity and differentiation within PNG.

From my analysis of these DNA sequences, *An. punctulatus, An. koliensis, An. farauti s.s., An. hinesorum* and *An. farauti 4* are supported as genetically independent species. *An. farauti s.s.* and *An. hinesorum*, the two most common and widespread
species, emerged as sister taxa but further ancestral relationships remain unclear. Phylogeographic patterns were observed suggesting geographically isolated, genetically distinct populations exist within species. Geographical features such as the Central Highlands and Bismarck Sea are possibly contributing to these observations of phylogeography. This sequence analysis enabled the development of genetic diagnostic tools that demonstrated proposed genetic species definitions are consistently observed across the country.

By identifying these five species as genetically independent and harboring geographically distinct subpopulations, it is conceivable that each species/subpopulation might exhibit different characteristics making them more/less efficient vectors of disease. This observed independence signifies the necessity to examine characteristics (e.g. ecology, biting habits, larval habitat, insecticide susceptibility, vector capacity) of each species and subpopulation to determine their importance as a vector within a community. The molecular tools and genetic information presented here will aid future studies and may improve interventions targeting specific vector species/populations for sustainable management of malaria and filariasis in PNG. Finally, whole-genome sequencing of these mosquitoes proposes to further refine understandings of species relationships as well as evolutionary origins of the species group.
I. CHAPTER 1

Anopheles punctulatus species complex:
The Status of Disease Vectors in Papua New Guinea
A. Introduction: Malaria transmission and the Anopheles vector

*Plasmodium* parasites are the causative agents of malaria disease and are transmitted to their human hosts in tropical and subtropical regions of the world. According to the World Health Organization and Malaria Atlas Project, over half of the world’s population lives in regions at risk for malaria transmission (Hay et al. 2009, Guerra et al. 2010, WHO 2010). Sir Ronald Ross first implicated mosquitoes as the vector of malaria parasites in 1898. Ross observed a cohort of healthy sparrows after being bitten by ‘grey mosquitoes’ that had previously fed on diseased sparrows—presumably infected with an avian *Plasmodium* species. Seventy-five percent of the previously healthy sparrows developed an infection with high parasitemia within a week’s time (Ross 1901). It is now known that *Plasmodium* parasites causing human malaria (*Plasmodium falciparum, P. vivax, P. malariae, P. ovale*) rely solely on mosquitoes from the genus *Anopheles* as their vector hosts. Female *Anopheles* mosquitoes are
considered the definitive host for the *Plasmodium* as sexual reproduction of the parasite occurs in the mosquito gut. In the process of taking a blood meal to obtain nutrients needed for egg production, the infectious female mosquito injects infectious *Plasmodium* sporozoites into the human [Figure 1.1 (White 2004)] (Service 2004, Mullen and Durden 2009).

Other *Plasmodium* species are known to cause malaria in birds, reptiles, rodents and non-human primates. A comprehensive list of rodent and non-human primate primary malaria vectors is somewhat incomplete, however of the known vectors, all are *Anopheles*. Those currently unknown are presumably also *Anopheles* as *Anopheles* feed primarily on mammals (Service 2002, Mullen and Durden 2009). However, *Anopheles* mosquitoes are not the primary vectors of avian and reptilian malaria. Members of the Genus *Culex* (Figure 1.2) are implicated as the primary avian malaria vectors and *Culex* species, as well as biting midges and sandflies, are the primary vectors of reptilian malaria (Mullen and Durden 2009).

Blood-feeding behavior of female *Anopheles* are known to vary in a species-specific, and sometimes strain or location-specific manner (Benet et al. 2004b, Service 2004). These behaviors can be broken down into specific categories. Blood-source preferences are shown in variable degrees with *Anopheles* species, where some species almost exclusively prefer humans (anthropophagic) and others prefer warm-blooded animals other than humans (zoophagic). This trait is not definitive and is usually influenced by local availability of hosts. Almost all blood-feeding by *Anopheles* occurs between dusk and dawn, however peak biting times may differ between species. Some species feed primarily outdoors (exophagic) while others
feed indoors (endophagic). Additionally, after taking a blood meal the mosquito must rest to digest and eliminate excess liquid. Where this resting takes place has also been shown to vary, as some species/strains prefer to rest on the walls inside a dwelling (endophilic), where others will leave the dwelling to rest in specific outdoor locations (exophilic) (Service 2002). Examining evolutionary history of Anopheles mosquitoes may aid in the understanding the relationships between Anopheles feeding behaviors and human malaria transmission.

The genus of Anopheles consists of six subgenera (Figure 1.2). Four subgenera, Kertezia, Lophopodomia, Nyssorhynchus, Stethomyia are restricted to South America, make up only 11% of Anopheles species and contain very few human disease vectors. Cellia (found in Africa, Asia, Europe and Australia) and Anopheles (cosmopolitan) represent 48% and 41%, respectively, of the species in the Anopheles genus (Service 2002, Krzywinski and Besansky 2003, Harbach 2004). Formation of these subgenera lineages has been dated to 95 million years ago (MYA), prior to the breakup of western Gondwana.

Given the absence of Cellia species in North
and South America, it is suggested that the time period this subgenus appeared can be forward-dated as occurring in the Eocene period (56-34 MYA) after South America broke off Gondwana (~30 MYA). It has also been suggested that the radiation and spread of subgenera Cellia and Anopheles could have been triggered by a radiation of their mammalian hosts. Given their greater species numbers as well as broad and tropical distribution, Cellia and Anopheles subgenera also harbor most of the malaria vector species (Krzywinski and Besansky 2003).

Of the approximately 500 identified species of the genus Anopheles, 70 have been shown to be competent vectors of malaria (Service 2002, Harbach 2004, Kiszewski et al. 2004, Hay et al. 2010). Upon close genetic inspection it is observed that a majority of these species cluster within groups of closely related, morphologically indistinguishable species complexes. Much of the diversity within these species complexes is poorly characterized and the vectorial capacity of the many sibling species is not understood (Service 2002, Harbach 2004). The most notorious malaria vector and highly anthropophilic species, Anopheles gambiae, has been most extensively studied.

B. The Australo-Papuan Region

1. Geological and Zoological History

The Independent State of Papua New Guinea (PNG) comprises the eastern half New Guinea island, located north of Queensland, Australia, as well as surrounding island provinces and small coastal islands (Figure 1.3). PNG and Australia are situated on the Sahul continental shelf also known as the Australian
plate. New Guinea and Australia were connected as part of Gondwana, attached to part of what are today Antarctica (along the southern boarder of Australia) and the Indian subcontinent (along the western boarder). During the Cenozoic period (45 MYA) the Sahul continental shelf broke away moving north and east. At this time, what is now Australia and New Guinea as well as Antarctica was covered in rainforest. Around 15 MYA the northern movement of the Sahul plate eventually resulted in a collision with Sunda (Southeast Asian Plate). At approximately the same time the circumpolar ocean current was established causing the formation the
Antarctic polar ice sheet and causing much of central Australia to dry out to its current state of aridity, however New Guinea maintained the rainforest climate (Veevers 1991, Johnson 2009).

Following the Sahul/Sunda collision, additional smaller, southward moving tectonic plates along the Australian and Pacific Plate boundary have collided with New Guinea near its northern coast over the last 5 million years. These collisions have contributed to the extreme ranges of altitude observed on the island today, specifically the formation of the Central and Northern Coastal Ranges (Hamilton 1979, Veevers 1991, Polhemus and Polhemus 1998), and continue to contribute to the volcanic activity and earthquakes experienced in Indonesia and PNG today (Johnson 2009).

Australia and New Guinea harbor a unique zoological composition that is unlike the flora and faunal characteristics of neighboring Southeast Asia. The floral and faunal patterns of this Australo-Papuan zoogeographic region are thought to have been influenced by the historical position of Sahul in an isolated corner of Gondwana, with no connection to Asia. In the past 15 million years, following the collision of the Sahul and Sunda (Figure 1.4), the Australo-Papuan and Asian species have colonized, adapted, and evolved across the Malay Archipelago in both eastern and western directions. During intervening glaciation periods where sea levels have been thought to vary as much as 150 meters, Australia, New Guinea and many smaller western islands were connected creating greater opportunity for species dispersal. Even though 15 million years have passed with Sunda and Sahul in proximity to one another their historical disconnect is still very much apparent in
both vegetation and animal life. This separation of Australo-Papuan and Asian fauna has been demarcated by many hypothetical boundaries including the Wallace, Lydekker and Weber lines separating islands of the Malay Archipelago (to the west) with >50% Asian fauna from islands rimming Sahul (to the east) with >50% Australo-Papuan fauna (Figure 1.4) (Mayr 1944). This faunal divide has been shown to extend to the Insect Class (Gressitt 1956, Holloway 1998, Beck et al. 2006, Balke et al. 2009) and may well have an influence on mosquito species dispersal.

Figure 1.4 Figure adapted from Mayr (1944) demonstrating location of Wallace, Weber and Lydekker Lines ('Limit of Australo-Papuan Mainland Fauna') of faunal demarcation in the Malay Archipelago. As well as the Sunda (Asia-Borneo) and Sahul (Australia-New Guinea) continental shelves.
2. **Geography and Climate**

Papua New Guinea’s total landmass of 462,840km\(^2\) encompasses very distinct and diverse geographical features (Figure 1.5). The Central Highlands, steep and rugged mountain ranges that characterize the center of the country, extend from Indonesian West Papua (formerly Irian Jaya) to the west, through the eastern tip of the mainland. These mountain chains, Muller, Central and Owen Stanley Ranges, reach as high as 4500m and include highland valley regions (1500-3000m) (McAlpine 1983, Allen 1992). The northern coast of the mainland is also characterized by a smaller mountain range, the Northern Costal Ranges, which are composed of the Torricelli Mountains and Prince Alexander Range along the northwestern coast, Albert Range along the north central coast and the Finisterre and Saruwaged Ranges along the northeastern coast reaching elevations of 900-1500m. The island provinces of New Britain (East and West), New Ireland, and Bougainville are also characterized by central mountain ranges reaching 1200m. These large geographical barriers influence diverse weather patterns experienced by regions of the country (McAlpine 1983) and contribute to extensive habitat diversity. Geography and climate characteristics, specifically temperature and rainfall/humidity, are key determinants of mosquito species range (Service 2004).

The country can be generally divided into three regions: north coast, Central Ranges and southern plains. Each of these general regions experiences different climatic patterns. The southern plains region is described as monsoonal and experiences the least amount of rainfall throughout the year with less than 2000
mm annually in regions south of the dashed line and 2000-3000 mm annually in regions north of the dashed line in Figure 1.5. Other than the Sepik River region which experiences less than 2000 mm of rainfall, the remainder of the mainland, and islands experience 3000-4000 mm rainfall annually with a maximum of 9000 mm in isolated regions of the Central Ranges (McAlpine 1983).

Figure 1.5 Papua New Guinea general maximum elevation and main geographic features.

The geographical characteristics of PNG combined with rainfall appear to have influenced Anopheles distribution. For example, the Central Ranges receive the highest amount of rainfall however the high altitude brings cooler temperatures to which a majority of the Anopheles have not adapted. Therefore abundance and distribution of malaria vectors in the highlands is minimal and limited to species that are thought to have adapted to these cooler conditions (Cooper et al. 2002). The Central Ranges also pose a geographical barrier for lowland species dispersal.
from one side of the ranges to the other. This is observed where some species are restricted to one side of the Central Ranges. The drier monsoonal southern plains of the Fly River region, where the temperature is similar to all other low-lying costal regions, receives much less rainfall creating the characteristic dry plain savannah that is also found in northern Australia. This hot, drier region also poses a unique habitat type also to which not all *Anopheles* have adapted.

C. *Anopheles* in Papua New Guinea

1. Historical Entomological Accounts

Members of the *Anopheles punctulatus* species complex were first described in the early 1900s-- *Anopheles punctulatus* (Dönitz 1901), *An. farauti* (Laveran 1902) and later *An. koliensis* (Owen 1945, Rozeboom and Knight 1946). These species belong to the subgenus *Cellia* (Neomyzomyia series) (Figure 1.2 (Besansky and Fahey 1997, Krzywinski et al. 2001b, Krzywinski and Besansky 2003)) proposed to be distinctly different from all other Asian and Australasian Neomyzomyia species groups through careful examination of adult morphological features. Specifically, these features included the presence of scales on the dorsal region of the thorax, absence of scales from all abdominal segments except terminal segments, speckling of the legs, and finally broad light apical bands on the last two palpal segments (Belkin et al. 1945). In 1923, George Heydon identified sporozoites and oocysts in anophelines in Rabaul (Heydon 1923).

For much of the 20th century the species complex was thought to be limited to these three species (*An. punctulatus, An. koliensis* and *An. farauti* s.l.).
range of the AP complex extends from the Moluccan Islands (Spice Islands) east through New Guinea and into the Solomon Islands, with *An. farauti* common in northern Australia (Rozeboom and Knight 1946). Species identification by morphology rests largely on the presence or absence of a sector spot along the costal wing vein and coloration of the proboscis (Figure 1.6) (Belkin 1962).

Further cross-mating and genetic studies from the 1970s to 2008 demonstrated that the species complex was much larger than the initial three members: *An. farauti, An. punctulatus* and *An. koliensis*. In 1973, Joan Bryan reported on results of cross-mating experiments using progeny from females captured from 12 locations around New Guinea and northern Queensland, Australia, which had been identified by proboscis morphology as *An. punctulatus, An. farauti* s.l. or *An. koliensis*. Bryan et al.’s studies found that the colonies of mosquitoes fell into four distinct groups. When members of the same colony were mated their progeny were fertile, however when cross-mated between colonies their progeny were sterile.

![Figure 1.6 Morphological characteristics used for species differentiation among *An. punctulatus, An. farauti* s.l. and *An. koliensis*. Presence of the sector spot (circled in red) distinguishes *An. punctulatus* and *An. farauti* s.l. from *An. koliensis*. Further discrimination is based on proboscis (directing red arrow) scale coloration, however *An. farauti* s.l. and *An. koliensis* have been shown to display polymorphic coloration of the proboscis with varying degrees of dark and white scaling. *An. farauti* s.l. members (7 species) are indistinguishable from each other by morphology. Photo credit The Walter Reed Biosystematics Unit.](image-url)
Morphologically, she concluded that she had colonies of *An. punctulatus, An. koliensis* and two separate, yet morphologically indistinguishable, colonies of *An. farauti*. When the two Farauti colonies were cross-mated their progeny were sterile. Therefore she determined that she had identified two separate species, aptly naming the first colony (established from Rabaul) *An. farauti* 1 (currently known as *An. farauti* s.s.), and the second colony (from a swampy northern Queensland location) *An. farauti* 2 (currently *An. hinesorum*). By similar experiments Bryan also generated support indicating that *An. koliensis* as a separate species as it had once been thought to be an intermediate of *An. farauti* and *An. punctulatus* (Bryan 1973b). Mahon and Miethke employed similar cross-mating methods and describe *An. farauti* 3 (currently *An. torresiensis*) as a distinct species from *An. farauti* s.s. and *An. hinesorum* in Queensland (Mahon and Miethke 1982).

Using allozyme analysis, Foley, Bryan and others, identified four additional members of *An. farauti* s.l. through observation of unique electrophoretic patterns: *An. farauti* 4 (near Madang, Madang Province), *An. farauti* 5 (near Goroka, Eastern Highlands Province), *An. farauti* 6 (near Tari, Southern Highlands Province) (Foley et al. 1993), *An. farauti* 7 (currently *An. irenicus*) (Guadalcanal, Solomon Islands) (Foley et al. 1994). Foley concluded that these additional four *An. farauti* members, even though morphologically indistinguishable, displayed distinctly different allozyme electrophoretic patterns from one another and from the *An. farauti* s.s., *An. hinesorum* and *An. torresiensis* used as control comparisons (Foley et al. 1993, Foley et al. 1994). A fifth species, *An. species* near *punctulatus* (Western Province) was identified to be morphologically similar to *An. punctulatus*, and like the previous
description of the additional *An. farauti* s.l. members, also demonstrated allozyme electrophoretic patterns not shared by any of the other *An. punctulatus* complex species (Foley et al. 1995).

*An. clowi*, described in 1946, has only been recaptured twice, however Cooper and colleagues describe its unique PCR-RFLP (polymerase chain reaction-restriction fragment length polymorphism) pattern, the gold standard molecular species identification using *MspI* restriction enzyme to describe differences in the internal transcribed spacer unit 2 of rDNA between species, from a collection around Madang in 2000 (Rozeboom and Knight 1946, Cooper et al. 2000). *An rennelliensis* was thought to be a member of the *An. punctulatus* complex based upon morphological descriptions, however this species has not been captured for cross-mating studies or molecular analysis (Taylor and Maffi 1971, Beebe and Cooper 2000). The most recently described member of the AP complex, *An. farauti* 8, also morphologically identical to the other 7 members of the Farauti clade, was identified in Central Province through DNA sequence analysis (Bower et al. 2008) and has not been further validated.

Of the 12 known *An. punctulatus* species complex members described above, 9 have been found within PNG on more than one occasion. Four members (*An. punctulatus*, *An. koliensis*, *An. farauti* s.s. and *An. hinesorum*) are common and widespread throughout the country. *An. farauti* 4 is common and widespread in regions north of the Central Ranges. Five species were previously collected in PNG however they are rare in abundance and limited to small and specific habitat ranges
within PNG:  *An. torriensis* (southern plains of the Fly River region of Western Province), *An. farauti* 6 (highland river valleys of Southern Highlands and Enga Provinces), *An.* species near *punctulatus* (lowland river valleys north and south of the Central Ranges) and *An. clowi* (east of Madang town, Madang Province) (Foley et al. 1995, Cooper et al. 2002).

Other *Anopheles* species, not belonging to the Punctulatus complex, are also found within PNG but in much less abundance (Steffan 1966, Cooper et al. 2009a). *An. longirostris* (*Cellia*: Neomyzomyia Series), *An. bancroftii* (*Anopheles*: Myzorhynchus Series), *An. subpictus* (*Cellia*: Pyretophorus Series) and *An. karwari* (*Cellia*: Neocellia Series) are four of the more common non-*An. punctulatus* species complex members found in the region and are considered minor malaria vectors due to their patchy distribution, low abundance and infrequent observed infection with malaria parasites (Cooper et al. 2009a). Morphological taxonomy and genetic analysis of these species indicates *An. bancroftii*, *An. subpictus* and *An. karwari* belong to distinctly different clades within the *Anopheles* Genus from the AP complex. It is suggested that *An. longirostris* is the closest relative of the AP complex as it is also part of the Neomyzomyia Series (Foley et al. 1998, Harbach 2004).

2. **Species discrimination by molecular taxonomy**

With evidence of species differences, but well-known difficulties of species discrimination by morphometric analysis, molecular-based methods began to appear in the 1980s and 1990s. Banding patterns in polytene chromosomes were first utilized to differentiate, *An. farauti* s.s., *An. hinesorum* and *An. torriensis
(Mahon 1983). A Southern blot method was developed utilizing $^{32}P$ labeled species-specific probes, derived from whole genomic DNA from established colonies of *An. punctulatus*, *An. farauti* s.s., *An. farauti* 4, *An. farauti* 5, *An. farauti* 6, *An. koliensis*. This squash blot technique required mosquito abdomens to be ‘squashed’ onto a nylon membrane where they were then probed for presence of species-specific DNAs (Beebe et al. 1994). Allozyme electrophoretic analysis was also utilized to discriminate between members of this species complex (Foley et al. 1993). These methods enabled identification of AP complex members with a greater degree of accuracy than morphology.

In the early 1990s post-PCR (polymerase chain reaction) approaches emerged for *An. gambiae* (Paskewitz et al. 1993), as well as North American anophelines (Porter and Collins 1991), and were typically found to be more accurate and less labor intensive, especially in field settings, than other molecular tools available at the time (Taylor et al. 1993). DNA sequence analysis of ribosomal DNA (rDNA) and development of these techniques emerged from comparative studies of closely related eukaryotic species. Specifically, within the Order Diptera, two groups described the utility of the internal transcribe spacer (ITS) (Wesson et al. 1992, Schlotterer et al. 1994) as a useful DNA sequence for phylogenetic analysis and cryptic species differentiation (Collins and Paskewitz 1996).

Eukaryotic cells are observed to contain hundreds of copies of rDNA genes. Internal transcribed spacers 1 (ITS1) and 2 (ITS2) of eukaryotes separate the 18S - 5.8S, and 5.8S - 28S rDNA, respectively. Their function is unclear but they are
thought to play a structural role in proper pre-RNA folding before they are cleaved during rRNA maturation (Musters et al. 1990, Joseph et al. 1999). The DNA sequence of the ITS2 region has been shown to be particularly useful in distinguishing closely related species. Given that rDNA genes are copied in abundance within a cell and the ITS regions are flanked by highly conserved ribosomal coding regions, the ITS2 region has become a very attractive target for PCR based methods differentiating cryptic anopheline species (Paskewitz et al. 1993, Beebe and Saul 1995, Collins and Paskewitz 1996, Cornel et al. 1996, Cornel and Collins 1996, Malafrente et al. 1999, Onyabe and Conn 1999, Hackett et al. 2000, Beebe et al. 2001, Manonmani et al. 2001, Garros et al. 2004).

![Figure 1.7](image)

Figure 1.7 ITS2 PCR-RFLP patterns observed for 10 An. punctulatus complex members. Adapted from Beebe & Saul 1995 (Fig. 1). AF s.s.= An. farauti s.s., AH= An. hinesorum (formerly AF2), AT= An. torresiensis, AF4-AF6= An. farauti 4- An. farauti , AI= An. iredicus (formerly AF7) AK= An. koliensis, AP=An. punctulatus, AnP=Anopheles species near punctulatus.

Based on ITS2 species identification methods, Beebe and Saul developed a PCR-RFLP (restriction fragment length polymorphism) assay using the ITS2 gene to differentiate 10 members of the An. punctulatus species complex (Figure 1.7) (Beebe...
and Saul 1995). Through DNA sequencing, large insertions, deletions and polymorphic motifs were shown to differentiate members of the *An. punctulatus* complex (Beebe et al. 1999a). The ITS2 PCR-RFLP method proved to be more accurate than species identification by morphology alone. Additionally, they demonstrated the assay’s robust capability to identify the same species from geographically distinct regions. This was the first relatively quick and most accurate method of differentiating members of this species complex (Beebe and Saul 1995). Focus on ITS2 has provided a starting point for strain identification, investigation of habitat preferences, and evaluation of vector relationships (Beebe et al. 1999b, Cooper et al. 2002).

3. **Species distribution and larval habitat preferences**

The largest study, both in number of mosquitoes captured, and degree of coverage, to date on the *An. punctulatus* complex was conducted by
Cooper and colleagues from 1992-1998. This study systematically surveyed the entire mainland of PNG using helicopter and four-wheel drive vehicles over these seven years to collect *Anopheles* (n=22,970) from 795 sites. Larval through adult lifecycle stages were collected using a variety of methods in each locale (Figure 1.8). Members of the *An. punctulatus* species complex were found in 735/795 locations (Cooper et al. 2002).

From this survey generalities about individual species ecology were more clearly identified. Oviposition sites included coastal brackish swamps and naturally occurring pools of water varying in salinity. A smaller amount of larvae were also collected from artificial sites created by humans (wheel ruts, drains, wells etc.). CO$_2$ baited light traps were used to collect adult females. This collection data along with night human landing catch methods were then used to establish each species (adult) association with humans by comparing the amount of a specific species that was either

![Figure 1.9 General distribution of the five most common AP complex species adapted from Cooper (2002) Figures 3-7. Generally, *An. farauti* s.s. is found with in 1km along the entire coast. Further inland locations are demonstrated by the red and yellow groups.](image-url)
captured in/near human habitation versus uninhabited locations. Figure 1.9 summarizes the habitat ranges for the five most abundant species; *An. farauti* s.s., *An. hinesorum*, *An. farauti* 4, *An. punctulatus*, and *An. koliensis* (Cooper et al. 2002).

Since these five species are the most abundant and are implicated as important vectors of disease, I will focus on highlighting their larval habitats and distribution as described by Cooper (2002) through the 735 mosquito collection sites below.

**a. *An. farauti* s.s.**

*An. farauti* s.s. was collected at 239 locations all along the coast of PNG (>50% from within 1 km of the coast & >75% from within 10 km of the coast). *An. farauti* s.s. was the dominant species in these locations. *An. farauti* s.s. had a positive association with human habitation (p<0.001). Cooper suggests this relationship was observed because human habitation in villages around the coast is found in close proximity to freshwater sources that could serve as breeding habitats for this species. This species is thought to be tolerant to saline conditions but freshwater conditions are preferred for oviposition (Cooper et al. 2002).

**b. *An. hinesorum***

*An. hinesorum* was collected at 324 locations and the most widely distributed species in terms of distance from the coast. A majority of *An. hinesorum* (formerly *An. farauti* 2) was collected from south of the central ranges, where Cooper indicated it was the most common and widely distributed member of the species complex. *An. hinesorum* was also commonly collected north of the Central Ranges and usually collected in these areas along with other species (*An. farauti* 4 and *An.
**punctulatus**). *An. hinesorum* showed preference to lowland river valleys 10-100 km from the coast where oviposition locations were commonly natural sites like small ground pools and pools formed along riverbeds. This species was also positively correlated with areas of human habitation (p<0.001).

c. *An. farauti 4*

*An. farauti 4* was collected least often out of these 5 predominant species within the AP complex, collected from 43 locations north of the Central Ranges in low-lying river valleys and only found near the coast around Madang town (Figure 1.3 Madang Province, location ‘C’) and Lae (Figure 1.3 Morobe Province, location ‘E’), both larger coastal towns. This species was the dominant species collected from the Markham, Ramu and Sepik River valleys and predominantly used artificial breeding sites such as manmade drains, wheel tracks and pig wallows. This species was also found in close association with human habitation (p<0.001). In Lae >90% of the human landing catch was *An. farauti 4*. Interestingly this observation was not reproducible in areas around Madang, with very few *An. farauti 4* collected in human landing catch (quantity not reported).

d. *An. punctulatus*

*An. punctulatus* was collected from 21 locations throughout the country, except in areas of Western Province where the climate is drier and monsoonal (especially south of the Fly River). Infrequency of *An. punctulatus* collections was attributed to the potential *An. punctulatus* avoidance of CO₂ light traps; however *An. punctulatus* larvae were readily collected. The distribution of *An. punctulatus* was similar to that
of *An. hinesorum* where it is found in low-lying inland river valleys but was also found near and at the coast. *An. punctulatus* has notably adapted to artificial oviposition sites created by humans and almost exclusively breeds in wheel ruts and ditches created by human travel. An association with *An. punctulatus* and human habitation was not detected. Cooper and colleagues attribute this also to the infrequent collection of *An. punctulatus* by CO$_2$ light traps (Cooper et al. 2002).

e. *An. koliensis*

*An. koliensis* was collected from 246 locations throughout the country. This species was found from the coast to 100 km inland with a majority of the collections occurring north of the central ranges in the Sepik and Markham River valleys. Additionally, this species was found south of the central ranges but in greater quantities along the eastern Papuan Gulf, and rarely in the Western Province region. Oviposition sites for *An. koliensis* were commonly a mixture of both natural (ground pools and watercourses) and artificial (wheel ruts and drains) and were correlated with human habitation (p<0.001).

f. Location of minor *An. punctulatus* complex members found in PNG

*An. torresiensis*, formerly *An. farauti* 3, was only collected from 13 locations and only in the drier Southern Plains/Fly River Region. This species is commonly found in Northern Territory, Australia where the climate is similar to that of the monsoonal PNG Southern Plains. *An. farauti* 5 was not collected in Cooper’s survey and has only been recorded once in the highlands (Foley et al. 1993, Foley et al. 1998) where it was first described. *An. farauti* 6 was collected from 8 locations and
only at high elevations (>1500 m) in river valleys and highland plains. *An.* species near *punctulatus* was collected from 10 locations and found in three locations in the low-lying Sepik River Valley as well as in a similar habitat south of the Central Ranges. Lastly *An. clowi* was only collected from 1 location, as larvae, to the east of Madang town (Cooper et al. 2002). *An. farauti* 7 (now formally known as *An. irenicus*) has only been reported in the Solomon Islands (Foley et al. 1994, Schmidt et al. 2003). *An. farauti* 8 has been collected in one Central Province (Figure 1.3) location (Bower et al. 2008).

4. **Understanding relationships of AP sibling species**

Relationships among the *Anopheles punctulatus* species complex are unclear. It is thought, based upon their indistinguishable morphology (dark, but polymorphic, proboscis and presence of the wing sector spot), that all members of the Farauti clade (*An. farauti* s.s., *An. hinesorum*, *An. torriensis*, *An. farauti* 4-6, *An. irenicus* and *An. farauti* 8) are closely related to each other. However many of these species display markedly different ecological preferences (i.e. *An. farauti* s.s. is found typically within 1 km of the ocean coast whereas *An. farauti* 6 is only found in high altitude intermountain valleys, >100 km from the coast) and some can be found sharing the same breeding habitat (i.e. *An. farauti* s.s., *An. hinesorum* and *An. farauti* 4) but no hybridization has been reported. The relationships between *An. punctulatus*, *An. koliensis* and the Farauti clade are puzzling. *An. punctulatus* shares the sector spot with the Farauti clade but does not have a dark proboscis. *An. koliensis* has been shown to have polymorphic dark proboscis coloration like that of
An. farauti s.l. but lacks the sector spot. Establishing relationships of the AP complex by morphology and ecology alone may be troublesome. Therefore further investigation of genetic relationships among the species complex would be necessary to understand genetic species definitions, ancestral relationships, shared behavioral and ecological traits among many other characteristics of this species complex.

What is currently known regarding phylogenetic relationships among the AP sibling species has been established using three different DNA sequences: mitochondrial cytochrome oxidase II (Foley et al. 1998), ITS2 (Beebe et al. 1999b) and 18S small subunit rDNA (Beebe et al. 2000d). These analyses were performed using one or very few individuals from each species as representative sequences. Therefore diversity within each species was not established. Results summarized in Figure 1.10 demonstrate these analyses.

Consistently, An. farauti s.s. and An. farauti 7 (An. irenicus) group together with high support (according to the 18S tree). Additionally, An. farauti 4, An. punctulatus and An. species near punctulatus are represented in the same clade across all three genes. The placement of An. koliensis in the phylogeny is different for each tree, however bootstrap support in all arrangements is not high. Only one representative from each species was used to construct these trees where variations of parsimony methods were used in a variety of programs.
Beebe and colleagues indicated that the arrangement of the 18S tree followed a general pattern that could be inferred back to morphology. Where An. farauti 4, An. punctulatus and An. species near punctulatus could be considered in their own clade—the Punctulatus clade—as they generally could be described as having a proboscis with the apical half colored white (with An. farauti 4 having a polymorphic proboscis coloration ranging from half white to entirely black). The Farauti clade was then made up of all the remaining Farauti members in the tree. These are thought to have an all black proboscis, although it has been reported to be polymorphic. An. koliensis was not included in either clade and could be considered
similar to *An. farauti* 4 in that the coloring of the proboscis is polymorphic (ranging from half white to entirely black) (Beebe and Cooper 2002).

Another observation of relationships represented by the phylogenetic trees is that of *An. farauti* s.s. and *An. farauti* 7. For field identification purposes, distinguishing *An. farauti* s.s. larvae from other members of the Farauti clade (Sweeney 1987) as well as *An. koliensis* and *An. punctulatus* (Foley et al. 1993) was frequently completed by a saline tolerance test (STT). In this method larvae are exposed to freshly collected seawater for one hour; those that survive are deemed tolerant (Foley and Bryan 2000a). Given that *An. farauti* s.s. can survive in both fresh and ocean water and most other members of the species complex (*An. hinesorum, An. torriensis, An. farauti* 5, *An. farauti* 6, *An. punctulatus* and *An. koliensis*) are thought to be sensitive to saline larval conditions (Sweeney 1987, Foley et al. 1993), STT was completed to differentiate *An. farauti* s.s. larvae from other AP complex members. In Guadalcanal only four members of the AP complex have been documented. *An. farauti* s.s. is the most abundant species found, and larvae of *An. punctulatus, An. hinesorum* and *An. farauti* 7 are commonly found with *An. farauti* s.s. in the same breeding sites especially during the dry season when breeding sites are not as abundant (Foley et al. 1994, Beebe et al. 2000b). Foley and Bryan have demonstrated that *An. farauti* 7 is also tolerant of saline larval conditions even though larvae have only been observed in freshwater pools (Foley and Bryan 2000a). They suggest that the STT is an unreliable method for distinguishing *An. farauti* s.s. larvae from *An. farauti* 7, because both species are morphologically indistinguishable and are tolerant to saline conditions (Foley and Bryan 2000a).
According to previously published phylogenetic reconstructions of the AP complex, *An. farauti* s.s. and *An. farauti* 7 are thought to be closely related (Figure 1.10) and it has been suggested that *An. farauti* 7 diverged from the ancestral *An. farauti* s.s. population on Guadalcanal (Hasan et al. 2008). This study revealed the saline tolerance trait is restricted to these two species within the phylogenetic reconstruction of known *An. punctulatus* sibling species.

D. *Anopheles* and *Plasmodium*

1. Human migrations and introduction of malaria into New Guinea

The Central Highlands have not only affected the microclimate of the country and species dispersal but also have played an important role in influencing settlement and distribution of the human population. It is thought that humans first colonized New Guinea 50,000 years ago, starting with the low-lying coastal regions. Archeological evidence suggests colonization of the less habitable central highlands occurred 45,000 years ago (Lilley 1992, O'Connell and Allen 2004, Summerhayes et al. 2010). The two major waves of early migration into New Guinea occurred before the end of the Pleistocene Epoch (12,000 years before present) and were associated with the colonization of Australia (populating mostly southern New Guinea) and the Proto-Papuan language group (populating parts of Western Papua and much of the highlands). These two distinct waves were then followed by a continuous colonization by people with maritime capabilities, most closely related those who have colonized the many Pacific islands to the east of New Guinea (Kirk 1992, Lilley 1992). The people of New Guinea were unknown to the European world until.
Spanish discovery in the 1500s. However, due to the rugged terrain of the Central Ranges, populations living in the Central Highlands were not discovered until the 1930s (Leahy 1936).

Aside from the rugged terrain, another factor that is thought to influence population is the presence of malaria. It is hypothesized that malaria was introduced into the Malay Archipelago ~5-10,000 years ago by the easterly migration of humans (Carter 2003). At this time, for the parasite to survive and establish endemicity, would have required a generally dense population of humans living in a location where competent Anopheles vectors were present. Presumably members of the An. punctulatus complex were present in New Guinea well before human colonization. However, archeological data establishing human population density of New Guinea has not been found and a formal census of Papua New Guinea population was not assessed until 1966 (Riley 1992). The United Nations estimated PNG's population to be 6.5 million in 2008 (UNdata 2010).

Today, malaria is generally endemic to regions below 1500 meters where the Anopheles vector is abundant (Peters 1960b, Peters 1960c). Below 1500 meters humans are exposed to malaria at very young ages allowing for the development of a degree of immunity. Due to lack of malaria exposure, those living in the highlands are not afforded the opportunity to develop immunity, creating the occasion for malaria epidemics in non-immune highland populations (Peters 1960b). Therefore a distinct partition of the human population can be observed between lowland and coastal regions and those living in the highlands. This partitioning is thought to be a
result of a emigration of local populations in response to high malaria morbidity and mortality in mid-altitude regions (Parkinson 1974, Riley 1983).

2.  **Host feeding preference**

Host preference of the AP complex has not been clearly established. The most recent analysis of the Human Blood Index (HBI) for this species complex has only been described for *An. farauti* s.s., from a very specific region around Madang town. Charlwood et al (1985), collected resting *An. punctulatus* (n=19), *An. koliensis* (n=133) and *An. farauti* s.s. (n=6419) as well as a non-sibling anopheline, *An. longirostris* (n=271), from eight villages in rural Madang. Collections of *An. punctulatus* and *An. koliensis* were not large enough to derive a statistically significant HBI. As for the *An. farauti* s.s. which were collected in abundance, HBI ranged from 9% to 83% between villages with a mean of 49.5%. Eleven percent (703/6419) of the *An. farauti* mosquitoes collected were analyzed electrophoretically and determined to be *An. farauti* s.s. Because all *An. farauti* collection locations were near the coast, Charlwood and colleagues assumed the remainder were also *An. farauti* s.s. (Charlwood et al. 1985).

Charlwood’s results suggested that *An. longirostris* fed primarily on pigs even though ample other available blood sources existed in the villages, demonstrating that perhaps *An. longirostris* was not an important vector for human malaria in this area (Charlwood et al. 1985). Cooper and colleagues have observed variable biting patterns for *An. longirostris* in villages around Madang during 1995 surveys. They have reported that *An. longirostris* was the dominant Anopheline collected by
landing catch in two Madang villages, however in surrounding villages it was rarely if ever collected (Cooper et al. 2006). Together these results suggest that vectorial status of *An. longirostris* is unclear and may be site specific.

It has been observed that species within the AP complex exhibit different biting time behaviors, which can vary within species from different locations. Benet et al. (2004) described the biting time behaviors of *An. farauti s.s.*, *An. hinesorum* and *An. farauti* 4 in villages within 50 km of Madang (Figure 1.3 near location ‘C’) and the Wosera Region of East Sepik Province (Figure 1.3 near location ‘B’). They demonstrated that within both Madang and Wosera regions *An. farauti* s.s. displayed a consistent biting rate from 1800hr to 0500hr, whereas *An. hinesorum*‘s biting was more focused on 1800hr-2100hr and *An. farauti* 4 had a bimodal biting pattern with an initial peak at 2100hr and a second peak at 0200hr.

Of further interest, Benet et al. reported that *An. koliensis* was observed to be comprised of three different genotypes by ITS2 PCR-RFLP: ‘M’, ‘W’ and ‘MW’, where ‘M’ variant was found only in coastal Madang, ‘W’ was found only in the Wosera and ‘MW’ was predominant in the Wosera but also found in low frequencies in Madang. These genotypes varied with respect to their biting times. ‘M’ variant in Madang was found to exhibit later biting times (2400-0500hr) and ‘W’ variant in the Wosera was found to exhibit earlier biting times (peaks from 1800-2200hr). The ‘MW’ variant collected from both Madang and Wosera was found to exhibit a more consistent biting time from 1800-0200hr (Benet et al. 2004b).
3. Vectorial capacity

In PNG, the vectorial capacity of each sibling species within the *An. punctulatus* complex has not been thoroughly established. Directly testing the mosquito's competence to propagate a malaria infection has not been assessed through feeding studies. Evidence implicating members of the species complex as malaria vectors was first described in 1923. *An. punctulatus* from Rabaul (East New Britain Province) was observed to have parasite development in the gut and sporozoites in the salivary glands after feeding on a *Plasmodium*-infected child (Heydon 1923). Further studies in the 1950s-70s concluded that *An. punctulatus*, *An. koliensis* and *An. farauti* s.l. were the main vectors of malaria and filariasis due to their abundance and observed propensity for biting humans where these disease were present (Black 1955, Spencer 1974). In the 1980s Charlwood and others established localized estimates of *An. punctulatus*, *An. farauti* s.l. and *An. koliensis* human biting rates. Human blood index from *An. punctulatus* group members in a region along the Pacific Coast near Madang town (Figure 1.3) demonstrated the blood feeding behavior of the species complex was highly variable (from 9-83%) between villages in close proximity (1-10 km). However, in general, of the *An. punctulatus*, *An. koliensis* and *An. farauti* s.l. collected that had taken a blood meal, >50% were found to have fed on humans. Other predominant blood sources included, pigs, dogs, cats, opossum and chickens (Afifi et al. 1980, Charlwood et al. 1985, Charlwood and Alpers 1986).
Findings from Benet et al. described the vectorial status of members of the *An. punctulatus* complex in Madang (Madang Province) and the Wosera (East Sepik Province) (Benet et al. 2004b). A total of 583 AP complex members (morphologically identified as *An. farauti* s.l.) from the Madang region were analyzed for *Plasmodium falciparum* and *P. vivax* circumsporozoite (CSP) antigen. Thirteen of these mosquitoes (0.2%) were positive for one or both *Plasmodium* species. Additionally, these mosquitoes were further classified using the ITS2 PCR-RFLP and determined to be one of four different species: *An. farauti* s.s., *An. farauti* 4, *An. koliensis* and *An. punctulatus*.

Collections from the Wosera were only classified by morphological species identification methods (Figure 1.6) (Belkin 1962). *An. punctulatus* (n=5612) and *An. koliensis* (n=6437) predominated in these locations with *An. farauti* s.l. (n=467 presumably *An. hinesorum* based on location) in the minority. *An. punctulatus* demonstrated the greatest number of infected mosquitoes (1.5%), followed by *An. koliensis* (0.3%). *An. farauti* s.l. was not positive for *Plasmodium* infection by the *Plasmodium* CSP antigen ELISA. Small sample sizes could have led to the differences in *Plasmodium* infection rates between the species and between locations as both regions are thought to have the same rates of human malaria infection (Benet et al. 2004b).

Recently, Cooper and colleagues reported on *Plasmodium* infection in Anophelines of PNG from their 1992-1998 survey (Cooper et al. 2009a). Using a subset of collection locations (n=258) from their survey previously described
(Cooper et al. 2002), they analyzed 22,970 mosquitoes for *Plasmodium* circumsorozoite protein antigen. *An. punctulatus* was implicated as the most frequently infected AP complex species collected (1.22%), followed by *An. farauti* 4 (0.98%), *An. hinesorum* (0.84%), *An. koliensis* (0.46%) and *An. farauti* s.s. (0.40%). Interestingly, in contrast with previous reports (Charlwood et al. 1985, Cooper et al. 2006), 7.82% of captured *An. longirostris* (non-AP complex species) were positive for *Plasmodium*. The high number of *An. longirostris* infections was attributed to being collected in areas of very high malaria burden on the human population (Cooper et al. 2009a). These data suggest that members of the AP complex transmit malaria at relatively similar rates however local variation could occur, and is poorly understood. Moreover, further reports of *Plasmodium* surveys in the *An. punctulatus* complex have not been reported.

4. **Early Malaria Control in PNG**

Disease control efforts by way of controlling the vector *Anopheles* did not begin until the 1940s during the Pacific conflict of World War II. These efforts focused mainly on areas where Allied troops were stationed and neglected all other regions of the country (Gunther 1974, Parkinson 1974). Blanketing areas with DDT (dichlorodiphenyltrichloroethane) (mixed with oil) from aircraft as well as oiling free-standing water to smother larval habitats surrounding military populations were the main methods of vector control measures employed (Gunther 1974). Beginning in the late 1950s and continuing through the 1970s, IRS of DDT was conducted as part of the WHO-initiated Malaria Eradication Program and reached
approximately 50% PNG’s population (Peters 1960c, Avery 1974, Parkinson 1974, Spencer 1992). Some provinces (Milne Bay, Northern, Chimbu, East and Western New Britain, New Ireland, Manus, and Bougainville) were reported to have approached 100% coverage in 1973. However others, for example Madang Province only reached 4.3% population coverage. Even though the program was moderately successful in controlling malaria, especially in reducing periodic highland epidemics, the spray campaign ceased after 1973 due to operational failure (Parkinson 1974). In 1984, based on the inability of the Malaria Eradication Program spray campaigns to reduce malaria to eradication levels, the PNG National government recommended operations cease (PNGDoH 1986, Genton et al. 1994, Mueller et al. 2003) and malaria control responsibilities were shifted to individual provincial governments (PNGDoH 1986). IRS of DDT continued at an infrequent rate throughout the 1980s in some locations where it was supported by local governments and Non-Governmental Organizations (NGOs) (Opeskin 2009, PMISG 2010). However in other locations, for example around Madang (Figure 1.3, location ‘C’), IRS was poorly accepted by residents, yearly DDT spraying ceased after 1979 and no other vector control measures were reported in this region throughout the 1980s (Cattani et al. 1986).

In 1989, the National government began promoting the availability of low cost insecticide impregnated bednets (ITNs) based upon studies in the Wosera and Madang demonstrating the benefits of ITNs (Graves et al. 1987, PNGDoH 1991, Genton et al. 1995, Hii et al. 2001). However no formal large-scale ITN distribution was implemented (PNGDoH 1991). ITN distribution throughout the early 1990s
reached >80% coverage in some small village clusters where bednet studies were being conducted (Genton et al. 1994, Mueller et al. 2003) or development projects were occurring (Hii et al. 1997) however true coverage has been shown to vary markedly (Bockarie et al. 2002). Additionally, due to the remoteness of most PNG villages, re-treatment of the distributed ITNs with permethrin rarely occurred (Genton et al. 1994). With the recent introduction of LLINs (long-lasting insecticidal nets) to the vector management repertoire, the PNG National Department of Health, Rotary Against Malaria, Global Fund and the Papua New Guinea Institute of Medical Research (PNGIMR) have recently (2004-2009) completed a broad distribution of LLINs throughout PNG (Hetzel 2010, Keven 2010).

E. **Depth of An. punctulatus complex published work**

When work on the PNGIMR/CWRU collaborative vector project began in 2005, little was known about the accuracy and durability of species definitions within the *An. punctulatus* species complex. Even though groups had performed large-scale collections covering most of the country (Cooper et al. 2002, Cooper et al. 2009a), genetic relationships within individual species had been addressed primarily by post-PCR RFLP analysis; DNA sequence analysis comparison had been performed on very limited numbers of mosquitoes primarily between species and not within species.

During the time-frame of the current studies, plans for mass distribution of free LLINs supported by the United Nation’s Millennium Development Project (Sachs 2005) and the Global Fund (PNGDoH 2004) were developing in PNG. Given
these newly available resources and the historical reminder that the previous WHO malaria eradication campaign in the 1950s and 60s was unsuccessful in controlling vector mosquitoes, there is a growing need to develop different strategies to evaluate malaria transmission by the *An. punctulatus* sibling species in Papua New Guinea.

*An. gambiae*, the main African malaria vector, has attracted the most attention among those concentrating on malaria vector biology, control and public health. In contrast, my dissertation will focus on members of the *Anopheles punctulatus* species complex (12 known species), thought to be the main vectors of malaria and filarial parasites in Papua New Guinea. To demonstrate the disparity in vector research on members of the *An. punctulatus* species complex, I conducted a simple comparison of published literature available on *An. gambiae* compared with that available for *An. punctulatus* species complex, as well other malaria vectors from Africa, India and South Asia, using U.S. National Library of Science PubMed search engine (Search Phrase: *Anopheles* [specific species], Limits: Field > Title). Table 1.1 illustrates the broad body of literature found on *An. gambiae*. A PubMed

<table>
<thead>
<tr>
<th>Species Key Word</th>
<th>Location of vector</th>
<th>PubMed</th>
<th>Google Scholar</th>
<th>Earliest Date</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>An. gambiae</em></td>
<td>Africa</td>
<td>1361</td>
<td>2,780</td>
<td>1946</td>
</tr>
<tr>
<td><em>An. punctulatus</em></td>
<td><strong>New Guinea/Aust</strong></td>
<td>45</td>
<td>78</td>
<td>1946</td>
</tr>
<tr>
<td><em>An. funestus</em></td>
<td>Africa</td>
<td>142</td>
<td>324</td>
<td>1947</td>
</tr>
<tr>
<td><em>An. dirus</em></td>
<td>South Asia</td>
<td>98</td>
<td>216</td>
<td>1981</td>
</tr>
<tr>
<td><em>An. stephensi</em></td>
<td>India</td>
<td>497</td>
<td>1070</td>
<td>1954</td>
</tr>
</tbody>
</table>

*Species key word was used for search in article title only


2 http://scholar.google.com

3 According to PubMed Search
title search revealed 1361 unique peer-reviewed articles on *An. gambiae* dating back to 1946. Comparing this with the 45 peer-reviewed articles about *An. punctulatus*, also dating back to 1946. This suggests a 30-fold greater understanding of the ecology, evolution, species and strain characteristics, vector competence, host preference, vectorial capacity, molecular genetics, and species distribution etc. of *An. gambiae*. For this reason *An. gambiae* becomes a natural comparison group for any research conducted on malaria vectors. In this dissertation, I will use what is known for *An. gambiae* as a platform to shape the direction and description of the work completed with *An. punctulatus* species complex.

**F. Dissertation Aims**

Since the discovery that *Anopheles* mosquitoes act as vectors of *Plasmodium* and other parasitic diseases, vector control methods have been suggested and implemented to reduce and eliminate disease burden (Ross 1910). Intensive vector control began with the 1939 discovery of the effective insecticidal properties of DDT (dichlorodiphenyltrichloroethane) coinciding with the onset of World War II. In 1955, the World Health Organization (WHO) began a malaria eradication program aimed at interrupting transmission of the disease through IRS of DDT to eliminate potentially infected resting anophelines as well as mass anti-malarial drug distributions to kill the *Plasmodium* parasites in human hosts in (WHO 1956). Sustainability and growing DDT resistance, among many other complications, led the WHO to shift the goal of malaria eradication to malaria control in 1969 (WHO 1969, Bruce-Chwatt 1986, Mendis et al. 2009). Given these factors combined with increasing anti-malarial chloroquine resistance, a rise in malaria incidence was
reported across the world. However, out of 143 countries where malaria was endemic in 1955, 37 were deemed malaria-free by 1978—a majority of these were in sub-tropical and temperate Europe as well as North and South America (WHO 1969, Wernsdorfer 1980).

The need for DDT alternatives and growing interest in malaria elimination/eradication (WHO 2006b, Gates and Gates 2007) has paved the way for a more informed approach to vector management, integrated vector management (IVM). This approach calls for individualized vector management plans tailored to the specific vector ecology of a disease endemic region as well as collaborating with ongoing public health initiatives and maintaining rational and sustainable use of insecticides and management practices. For example, IVM incorporates less toxic insecticides in a range of interventions such as LLINs coupled with IRS of insecticides including DDT where appropriate (WHO 2004).

Reducing malaria to elimination levels, where zero locally acquired cases occur, will require strong molecular epidemiology methods. In order for malaria elimination campaigns to succeed these methods must be developed and employed to monitor every stage in the disease life history including but not limited to parasite burden and species composition in the human population, vector species composition and abundance, vector-parasite species interactions, and parasite and vector responses to anti-malaria/mosquito interventions.

The purpose of my dissertation is to gain a better understanding of the AP complex—to understand genetic diversity within and among sibling species, evaluate current species definitions and relate these findings to improve Integrated
Vector Management. To achieve this overall goal my dissertation has focused on the following Aims.

**Aim 1**- Test the durability and accuracy of current species definitions while establishing a baseline understanding of genetic diversity within and between vector species and populations. For this dissertation I have chosen to focus on 5 members of the *An. punctulatus* complex as they are the most common malaria vector species in PNG (Burkot et al. 1988, Cooper et al. 2002, Benet et al. 2004b, Cooper et al. 2009a). The current multiple species hypothesis was tested through the following objectives:

1. Establish a collection of *An. punctulatus* species complex members from locations in PNG to achieve a diverse and broad sampling coverage.
2. Identify, amplify and sequence candidate genes from these representatives.
3. Determine the genetic diversity within populations of the same species as well as differentiation between species to assess evolutionary history.

**Aim 2**- Determine if relationships within the *An. punctulatus* complex can be verified using phylogenetic techniques.

1. Subject nuclear and mitochondrial DNA sequence to phylogenetic analysis to assess the durability of the multiple species hypothesis across different genes.
2. Utilize several phylogenetic approaches in attempt to identify ancestral species relationships.
3. Apply dating calibration points to phylogenetic data to infer divergence times of the *An. punctulatus* sibling species.

**Aim 3**- Investigate how genetic and phylogenetic data generated from the previous objectives could be exploited for use in monitoring integrated vector management strategies. This part of my dissertation will test the following hypotheses:

1. If robust species definitions exist and genetic characteristics are consistent with the multiple species hypothesis it is possible that this genetic information can be used to develop more efficient and accurate molecular species diagnostic assays that can be applied to mosquitoes captured throughout Papua New Guinea?

2. Can genetic tools be developed and applied to survey for mutations associated with insecticide resistance, in a species-specific manner, as intense insecticide pressure (mass distribution of LLINs) is being applied to the country?
II. CHAPTER 2

Rationale for Study Design and Study Methods
A. Introduction

The formidable African malaria vector, *An. gambiae* s.s., is a member of the *An. gambiae* species complex consisting of at least 7 morphologically indistinguishable species (Coetzee et al. 2000). Within the *An. gambiae* species complex it has been observed that *An. gambiae* s.s. populations vary genetically across their habitat range. This genetic variation is observable at the chromosome (giving rise to two chromosomal forms, M and S form) and individual gene level. Genetic variation is thought to result from habitat preference differences. Specifically, the M form of *An. gambiae* prefers breeding sites that are more permanent and rich with nutrients and predators, whereas the S form prefers temporary water pools that are typically free of predators (Coluzzi et al. 1979, Lehmann and Diabate 2008, Lawniczak et al. 2010, Neafsey et al. 2010). Given this observed differential habitat adaptation of the major African malaria vector, which belongs to a complex sibling species group, it becomes necessary to understand if similar differential habitat adaptation occurs or has occurred within are the *An. punctulatus* species complex.

In order to investigate the *An. punctulatus* species complex in Papua New Guinea a study design was carried out using methods described in this chapter.
B. Mosquito Collections, Morphological Identification and Genomic DNA Extraction

1. Mosquito collections

The Entomology Unit of Papua New Guinea Institute of Medical Research (PNGIMR) collected mosquitoes throughout PNG as part of surveillance studies associated with distribution of LLINs (Hetzel 2009), insecticide susceptibility trials (Keven 2010), and lymphatic filariasis elimination programs (Thomsen et al. 2010) among other ongoing projects. Mosquitoes were collected through night landing catch methods, larva collections, and Center for Disease Control light traps as previously described (Bockarie et al. 2000, Hii et al. 2000). Collections of larvae were reared into adults in the laboratory to allow for adult morphological identification. Collections span 24 villages from seven provinces as well as a range of altitudes and distances from the ocean coast. All locations were defined by Global Position System (GPS) latitude and longitude (Table 2.1 and Figure 2.1).

Figure 2.1 Locations of mosquito collections (n=24) in Papua New Guinea spanning 7 provinces. Location details are provided in Table 2.1.
Table 2.1
Mosquito Collection Site Data

<table>
<thead>
<tr>
<th>Site</th>
<th>Province</th>
<th>Altitude</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Distance to coast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albulum</td>
<td>EAST SEPIK</td>
<td>355</td>
<td>-3.5609</td>
<td>142.67565</td>
<td>20</td>
</tr>
<tr>
<td>Dreikikir</td>
<td>EAST SEPIK</td>
<td>375</td>
<td>-3.5445</td>
<td>142.77123</td>
<td>20</td>
</tr>
<tr>
<td>Nale</td>
<td>EAST SEPIK</td>
<td>104</td>
<td>-3.7725</td>
<td>143.07056</td>
<td>30</td>
</tr>
<tr>
<td>Nanha</td>
<td>EAST SEPIK</td>
<td>407</td>
<td>-3.5435</td>
<td>142.72650</td>
<td>20</td>
</tr>
<tr>
<td>Nghambule</td>
<td>EAST SEPIK</td>
<td>360</td>
<td>-3.5950</td>
<td>142.75000</td>
<td>20</td>
</tr>
<tr>
<td>Peneng</td>
<td>EAST SEPIK</td>
<td>254</td>
<td>-3.5435</td>
<td>142.65018</td>
<td>20</td>
</tr>
<tr>
<td>Yawatong</td>
<td>EAST SEPIK</td>
<td>204</td>
<td>-3.5713</td>
<td>142.68988</td>
<td>20</td>
</tr>
<tr>
<td>Bilbil</td>
<td>MADANG</td>
<td>4</td>
<td>-5.2885</td>
<td>145.76371</td>
<td>&lt; 0.5</td>
</tr>
<tr>
<td>Dimer</td>
<td>MADANG</td>
<td>327</td>
<td>-4.8007</td>
<td>145.62395</td>
<td>7</td>
</tr>
<tr>
<td>Hudini</td>
<td>MADANG</td>
<td>27</td>
<td>-5.2890</td>
<td>145.74413</td>
<td>2</td>
</tr>
<tr>
<td>Nuru</td>
<td>MADANG</td>
<td>154</td>
<td>-5.4067</td>
<td>145.59528</td>
<td>15</td>
</tr>
<tr>
<td>Sausi</td>
<td>MADANG</td>
<td>160</td>
<td>-5.6999</td>
<td>145.53802</td>
<td>35</td>
</tr>
<tr>
<td>Yagaum</td>
<td>MADANG</td>
<td>180</td>
<td>-5.2904</td>
<td>145.79873</td>
<td>2</td>
</tr>
<tr>
<td>Lorengau</td>
<td>MANUS</td>
<td>68</td>
<td>-2.0348</td>
<td>147.26136</td>
<td>1</td>
</tr>
<tr>
<td>Gingala</td>
<td>MOROBE</td>
<td>4</td>
<td>-6.6304</td>
<td>147.86111</td>
<td>&lt; 0.5</td>
</tr>
<tr>
<td>Godowa</td>
<td>MOROBE</td>
<td>2</td>
<td>-6.5899</td>
<td>147.82233</td>
<td>&lt; 0.5</td>
</tr>
<tr>
<td>Mumeng</td>
<td>MOROBE</td>
<td>1119</td>
<td>-6.9314</td>
<td>146.60930</td>
<td>40</td>
</tr>
<tr>
<td>Ramu</td>
<td>MOROBE</td>
<td>444</td>
<td>-6.0900</td>
<td>145.98100</td>
<td>65</td>
</tr>
<tr>
<td>Pitapena</td>
<td>WEST SEPIK</td>
<td>183</td>
<td>-4.6297</td>
<td>141.10485</td>
<td>125</td>
</tr>
<tr>
<td>Yapsie</td>
<td>WEST SEPIK</td>
<td>185</td>
<td>-4.6278</td>
<td>141.09560</td>
<td>125</td>
</tr>
<tr>
<td>Kuru</td>
<td>WESTERN</td>
<td>43</td>
<td>-8.8718</td>
<td>143.06471</td>
<td>20</td>
</tr>
<tr>
<td>Tavakip</td>
<td>WESTERN</td>
<td>55</td>
<td>-8.7906</td>
<td>142.87154</td>
<td>25</td>
</tr>
<tr>
<td>Brimba</td>
<td>WHP*</td>
<td>1484</td>
<td>-5.5883</td>
<td>144.68255</td>
<td>120</td>
</tr>
<tr>
<td>Singoropa</td>
<td>WHP*</td>
<td>1357</td>
<td>-5.5420</td>
<td>144.65368</td>
<td>120</td>
</tr>
</tbody>
</table>

*WHP* Western Highlands Province, † Altitude in meters, ‡ Approximate distance in kilometers
§ Locations of insecticide susceptibility trials

Figure 2.2 Individual mosquitoes collected and analyzed for this dissertation. The purple boxes represent the sample population utilized for sequencing ITS2, VGSC, COI and Cytb (n=111) and evaluation (n=425) of one or both diagnostic assays (to be described: ITS2 LDR-FMA and/or VGSC LDR-FMA). * The dashed arrow indicates that 75 mosquitoes were subjected to one or both LDR-FMAs as well as one or more DNA sequencing reactions and reported in this dissertation. The remaining 35 sequenced mosquitoes were previously subjected to LDR-FMAs with results not reported in this dissertation. The blue box represents a field survey of 797 mosquitoes which Anopheles species diagnostic and Wuchereria bancrofti parasite detection (ITS2 +Wb LDR-FMA) were completed.
A total of 1416 individual female mosquitoes were collected and utilized in this dissertation for DNA sequence analysis and diagnostic assay development and validation. Figure 2.2 describes the extent to which portions of these mosquitoes were utilized for DNA sequencing, included for diagnostic assay analysis and/or both.


Morphological identification of female mosquitoes were completed using methods described by Belkin (Figure 1.6 & Figure 2.3) (Belkin 1962). Morphological identification allows for members of the An. punctulatus group to be classified as An. punctulatus, An. koliensis, or An. farauti s.l. (sensu lato); differentiating the eight members of the Farauti clade as well as An. punctulatus and An. sp nr punctulatus is not possible by morphological methods. The presence of the sector spot along the coastal vein of the wing in An. punctulatus and An. farauti s.l.
allow differentiation from *An. koliensis* (which lacks the sector spot as shown in Figures 1.6 and 2.3). Further differentiation of *An. farauti* s.l. from *An. punctulatus* is then determined by the coloration of the proboscis, where *An. punctulatus* is thought to have a proboscis characterized by light brown scales, *An. farauti* s.l. has much darker brown scaling with a patch of white scaling towards the tip. Both *An. farauti* s.l. and *An. koliensis* have been shown to demonstrate polymorphic degrees of dark and white scaling (Woodhill 1946, Belkin 1962, Foley et al. 1993, Beebe et al. 2000d). During the collection process mosquitoes morphologically identified as members of the *An. punctulatus* group were kept in coded vials containing silica gel until DNA extraction could be completed.

3. **Insecticide susceptibility bio-assays**

Five locations in three disease endemic PNG provinces (identified by ‘§’ in Table 2.1) were chosen to assess the insecticide susceptibility of the *An. punctulatus* group. Each of these locations had a different history of insecticide use, varying in coverage and intensity. These locations were chosen to capture the variety of insecticide exposure types mosquito populations are likely to encounter; from historical DDT IRS exposure to high coverage LLIN use to intensive agricultural pyrethroid exposure. Larval collections from each location were reared into adulthood in the laboratory. Two- to five-day old females were tested for the knock down resistance phenotype using two WHO insecticide susceptibility assays (Keven 2010).
First, the ‘tube assay’ consists of a clear, plastic tube with a filter paper barrier on one end. The filter paper is treated with 0.05% lambda-cyhalothrin to mimic conditions of indoor residual spraying. The representative female mosquitoes are exposed to the treatment for 60 minutes and removed from the treatment tube and observed. Secondly, the ‘cone assay’ was performed using a new 55 mg/m² deltamethrin-treated LLIN (produced by PermaNet®) and naïve (not tested in the tube assay) female mosquitoes. The net was attached to a surface angled at 45 degrees. Five clear plastic cones were attached to the net and 3-5 female mosquitoes were added to each cone using a manual aspirator through the open tip of the cone and blocked from exiting the cone by cotton. After a 3-minute exposure to the net, mosquitoes were removed and observed. For both experiments, post-exposure observations were taken at 1 hour to assess for knockdown resistance phenotypes and at 24 hours to determine 24-hour mortality rates. Also, for both the tube and cone assays, control experiments were conducted using untreated filter paper and untreated bednet, respectively, with female mosquitoes from the same cohorts (WHO 1998, 2005, 2006a). After phenotype analysis, mosquitoes were morphologically identified and DNA extraction was completed for further genetic analyses (Keven 2010).

4. **Genomic DNA extraction**

Genomic DNA was extracted from single whole female mosquitoes using either a Qiagen method (QIAamp 96 or individual spin blood and tissue kit—QIAGEN, Valencia, CA; recommended protocol) or by a modified method (Bender et
al. 1983). In the modified method, individual whole mosquitoes were thoroughly
pulverized by vortexing each mosquito with a copper BB in a 1.5 ml microfuge tube
containing 100 μl of grinding buffer (0.1 M NaCl, 0.2 M sucrose, 0.1 M Tris HCl [pH
9.1-9.2], 0.05 M EDTA and 0.5% SDS). Samples were incubated at 65°C for 30
minutes; 8 M KOAc was then added (final concentration 1 M) to each tube and
incubated on ice for 30 minutes. Samples were centrifuged at 13,500 rpm (15
minutes) and the supernatant was transferred to a new sterile microfuge tube; 100
μl of 100% EtOH was added for ethanol precipitation of mosquito DNA. Tubes were
incubated at room temperature (5 minutes) and then centrifuged at 13,500 rpm (15
minutes). Supernatants were removed; 100 μl of ice cold 70% EtOH was added to
each sample, mixed and centrifuged at 13,500 rpm (5 minutes). Supernatants were
removed and sample tubes were allowed to dry overnight; precipitated DNA was
resuspended in 30-100 μl of sterile diH2O or TE buffer.

C. Gene Selection Rationale

The PNGIMR/CWRU collaborative vector research project began in 2005. At
this time information regarding genetic diversity within and among An. punctulatus
complex species was not abundant. What was known had been summarized in
work by Beebe and colleagues (Beebe and Saul 1995, Beebe et al. 2000c).

First, in the validation of an ITS2 PCR-RFLP species diagnostic assay, Beebe
and colleagues analyzed ITS2 PCR-RFLP patterns of An. farauti s.s. (n=4), An.
hinesorum (n=3), An. farauti 3 (n=2), An. koliensis (n=3), An. punctulatus (n=3) and
An. species near punctulatus (n=2) collected from geographically different locations.
They then demonstrated that the same ITS2 RFLP electrophoresis patterns within each species was observed even though they were collected from geographically different areas in PNG (Beebe and Saul 1995).

Second, with further investigation into populations of An. farauti s.s. from 30 coastal locations around mainland PNG as well as island provinces and Australia, Beebe and colleagues later demonstrated slight differences in ITS2 heteroduplex banding patterns. This suggested that 7 unique ecotypes existed within An. farauti s.s. and were geographically distributed (Northern Territory, Queensland, Southern Coast PNG, Northern Coast PNG, Rabaul, Guadalcanal and Vanuatu). These ecotypes were then further divided into 13 distinct genotypes through ITS2 DNA sequence analysis. No phenotypic differences were evaluated for these populations of An. farauti s.s. (Beebe et al. 2000c).

In order to assess questions of diversity both within and among species, I chose to focus on the five most abundant and widely distributed malaria and filariasis vector species (both country-wide and specifically in CWRU/PNGIMR collaborative study locations) within the An. punctulatus species complex members—An. punctulatus, An. koliensis, An. farauti s.s., An. hinesorum and An. farauti 4 (Beebe and Cooper 2002, Benet et al. 2004b, Cooper et al. 2009a). Two nuclear and two mitochondrial candidate genes were chosen for the purpose of assessing diversity and relationships among AP complex sibling species.
1. **Internal Transcribed Spacer 2 sequence**

The internal transcribed spacer 2 (ITS2), a nuclear sequence, was chosen as previous studies with Diptera, specifically *Anopheles*, have suggested that rDNA spacer regions are useful for distinguishing cryptic species and for phylogenetic analysis (Paskewitz et al. 1993, Collins and Paskewitz 1996, Cornel and Collins 1996, Beebe et al. 1999b). This region of the genome was also being used at the time of this study as the ‘gold standard’ molecular PCR-RFLP identification method for the *An. punctulatus* species complex (Beebe and Saul 1995) therefore some previous knowledge about the sequence of AP complex members was known.

2. **Voltage Gated Sodium Channel gene**

A portion of the voltage gated sodium channel gene (VGSC), a nuclear gene, was chosen as insecticide resistance in malaria vectors from Africa, Asia and parts of the South Pacific has been attributed to a single mutation in this gene (L1014F), known as kdr (knock down resistance) (Martinez-Torres D 1998, Syafruddin et al. 2010). Surrounding the location of this mutation are two polymorphic introns that have been shown to be useful in population studies of *An. gambiae* (Martinez-Torres D 1998, Etang et al. 2009). Investigating the exon region where the kdr mutation site is located as well as capturing surrounding polymorphic introns was thought to be useful for diversity purposes. This approach would also help to characterize the presence or absence of kdr in the AP complex since descriptions of this gene and regions surrounding kdr have not been reported for the AP complex. More detailed rational is provided in Chapter 4.
3. **Mitochondrial genes**

Two mitochondria genes were chosen to complement the two nuclear DNA sequences. Due to the nature in which mitochondria are inherited (maternal haploid lineage) and their lack of recombination, as well as their abundance in the eukaryotic cell, mitochondrial genes are a useful target for amplification and for genetic comparisons and detecting gene flow among closely related populations (Avise 1987). It has also been shown that the rate of evolution in the mitochondrial genome is 5-10 times higher than that of the nuclear genome (Simon et al. 1994, Hartl and Clark 1997). These unique features of mtDNA inheritance and polymorphism have potential to provide a different perspective on relationships within and among the AP complex members. Mitochondrial gene selections are described below.

**a. Cytochrome Oxidase I**

In 2004, the Consortium for the Barcode of Life (CBOL) was established to address the growing interest in having a standard gene sequenced for all life forms. From this group came the development of BOLD (Barcode of Life Datasystems)— a bioinformatics interface that serves to organize and store the CBOL gene of choice, Cytochrome Oxidase I (COI), from all living organisms to facilitate further research at an organism, population and community level (Ratnasingham and Hebert 2007). This region of the mitochondrial genome has been studied extensively in evolutionary analyses of both distant and closely related species, especially Diptera species (Lunt et al. 1996, de Brito et al. 2002, Dusfour et al. 2004, Foley et al. 2007),
but no sequence characteristics of \textit{COI} have been documented for the AP complex. Given this growing practical utility in \textit{COI}, this gene became a natural candidate for addressing questions of diversity and relationships within the AP complex.

\textbf{b. Cytochrome B}

An additional mitochondrial gene, Cytochrome B (\textit{Cytb}), was chosen for investigation as it has shown to complement population analysis of \textit{COI} as well as other nuclear sequences (i.e. ITS2) in other cryptic \textit{Anopheles} species groups (Besansky et al. 1997, Mukabayire et al. 1999, Krzywinski et al. 2001a, Dusfour et al. 2004). Theoretically, since no recombination occurs in the mtDNA both mitochondrial genes should provide the same information, however it is possible that some genes accumulate more polymorphism than others. Therefore a second mitochondrial gene will be useful in validating observations in \textit{COI}.

\textbf{D. Polymerase Chain Reaction (PCR) Methods}

PCR amplification was performed for all four of the previously described sequences. All PCR amplification reactions were carried out in a 25 μl volume utilizing one of two buffers (Buffer 1 [ITS2, \textit{COI}]: 67 mM Tris-HCl [pH 8.8], 6.7 mM MgSO4, 16.6 mM (NH₄)₂SO₄, 10 mM 2-mercaptoethanol, or Buffer 2 [\textit{VGSC}, \textit{Cytb}]: 50 mM KCl, 10 mM Tris-HCl [pH 8.3], 1.5 mM MgCl₂, 0.01% gelatin), 100 μM dATP, dGTP, dCTP, and dTTP, 2.5 units of thermostable DNA polymerase and 3 pmoles of the appropriate up- and downstream primers (Table 2.2) and 3 μl of appropriate genomic DNA sample. The thermocycling conditions are given in Table 2.2 and performed using a BioRad DNA Engine Tetrad 2 Peltier Thermal Cycler (Hercules,
CA). To evaluate overall amplification efficiency, PCR products were separated by electrophoresis on 2% agarose gels (1X TBE), stained with SYBR® Gold (Invitrogen, Carlsbad, CA), and visualized on a Storm 860 using ImageQuant 5.2 software (Molecular Dynamics, Sunnyvale, CA).

| DNA Target | Primer Sequence
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ITS2†</td>
<td>ITS2up 5′-TGTGAACCTGCAAGACAGCAT-3′&lt;br&gt;ITS2dn 5′-TATGCTTAAATCCAGGGGGT-3′&lt;br&gt;95°C 2 min (1x), 95°C 30 sec, 55°C 30 sec, 72°C 1 min (35x), 72°C 4 min (1x)&lt;br&gt;603-717 bp</td>
</tr>
<tr>
<td>VGSC†</td>
<td>aDip1up 5′-CGCCGACCTGATTTACTC-3′&lt;br&gt;cDip2dn 5′-TTCGCAAAACAGCTAAGA-3′&lt;br&gt;95°C 2 min (1x), 95°C 30 sec, 60°C 30 sec, 72°C 1.75 min (39x), 72°C 4 min (1x)&lt;br&gt;1305-1381 bp</td>
</tr>
<tr>
<td>COI†</td>
<td>UEA3up 5′-TATGATCCCAAGAAATATA-3′&lt;br&gt;UEA10dn 5′-TCAATGCAATATCGGATTATA-3′&lt;br&gt;95°C 5 min (1x), 95°C 30 sec, 45°C 30 sec, 72°C 1.5 min (35x), 72°C 4 min (1x)&lt;br&gt;1270 bp</td>
</tr>
<tr>
<td>Cytb†</td>
<td>cytbup 5′-GGAAATATCTTTTTAGGAGCAACAG-3′&lt;br&gt;cYtbdn 5′-ATTATGCTCCTTAGTATTAGAATTTG-3′&lt;br&gt;94°C 5 min (1x), 94°C 30 sec, 45°C 30 sec, 72°C 1.5 min (35x), 72°C 7 min (1x)&lt;br&gt;470 bp</td>
</tr>
</tbody>
</table>


E. Cloning and DNA Sequencing

TOPO TA cloning (Invitrogen, Carlsbad, CA) of amplified regions of interest was performed on individual mosquitoes. Blue-white screening was performed and several white colonies were picked from each cloning reaction to ensure proper capture of the DNA sequence of interest. In order to assess proper insertion of the PCR product, PCR amplification was completed using primers and protocols previously described for each gene. Colonies found to harbor the insertion of interest were then subjected to bidirectional plasmid sequencing (BigDye Termination Sanger sequencing [Beckman Coulter Genomics, Danvers, MA]) in a 96
well format. Each representative mosquito was sequenced at least twice, both forward and reverse, to ensure quality and reproducibility of sequence data.

F. General Design of Ligation Detection Reaction-Fluorescent Microsphere Assay

Post-PCR Ligation Detection Reactions (LDR)- Fluorescent Microsphere Assays (FMA) were designed to capture species-specific polymorphisms as well as

<table>
<thead>
<tr>
<th>Gene Target</th>
<th>Probes</th>
<th>Primer sequence†</th>
<th>FlexMAP™ microsphere‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFs.s. Classification</td>
<td>5′-tcaaatatcatctcatctatagcct GCC CCC AGM GA-3′</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>AFs.s. Reporter</td>
<td>/SPhos/TCT GGC CCA CAT GCA C/3Biotin/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AH Classification</td>
<td>5′-ctatcatctcatctatagcct TCG GCC CCC AAG CCA TG-3′</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>AH Reporter</td>
<td>/SPhos/CAC ACT GGC CCA CAA CC/3Biotin/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP Classification</td>
<td>5′-ctatactatcagtgacgc GCG GGG GGC T-3′</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>AP Reporter</td>
<td>/SPhos/AGG AGC CCC TGC AG/3Biotin/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF4 Classification</td>
<td>5′-ctatactatcagtgacgc CCG GCC GAC TT-3′</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>AK Classification</td>
<td>5′-ctatactatcagtgacgc CCG GCC GAC TT-3′</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>AF4 &amp; AK Reporter</td>
<td>/SPhos/GGCT GCT GTY YAC ACT AG/3Biotin/</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*AFs.s. = An. funestus s.s., AH = An. hinsicorum, AP = An. punctulatus, AFs = An. funestus f., and AK = An. koliensis
†Nucleotides in lower case (24 bases) represent TAG sequences added to the 5′ end of each species-specific LDR probe.
‡Single letter nucleotide codes: M=A or C, K=G or T, Y=T or C
§Luminescent microsphere sets are synthesized to exhibit unique fluorescence and each microsphere set is coupled to different anti-TAG sequences which are complementary to the species-specific TAG sequence primers.

Table 2.3

Ligation detection reaction probes for PNG Anopheles species* identification and mutation detection
mutations associated with insecticide resistance and details of these assays will be discussed in further chapters. Following is a basic description of LDR-FMA design and procedures.

Following PCR amplification, products were added to a multiplex ligation detection reaction (LDR) where species-specific (or mutation-specific) upstream classification probes (containing TAG sequences specific to Luminex® microsphere sets) ligate to downstream sequence reporter probes (3’-biotinylated) when appropriate target sequences are available (Figure 2.4A & B, Table 2.3). LDRs were conducted in a solution (15 μL) containing 20 mM Tris-HCL buffer, pH 7.6, 25 mM potassium acetate, 10 mM magnesium acetate, 1 mM NAD+, 10mM dithiothrietol, 0.1% Triton X-100, 10 nM (200 pmol) of each LDR probe (specific appropriate probes for each assay Table 2.3), 1 μL of each PCR product and 2 units of Taq DNA ligase (New England Biolabs, Beverly, MA). LDR reactions were initially heated at 95°C for one minute, followed by 32 thermal cycles of 95°C for 15 seconds and 58°C for 2 minutes.

Classification and reporter labeling of LDR products has been modified from the previously described protocol (McNamara 2006). Classification labeling of LDR products with Luminex Flex-MAP microspheres (Luminex Corporation, Austin, TX) containing anti-TAG oligos complementary to the TAG on the sequence-specific primers (Table 2.3 and Figure 2.4C) occurs in a hybridization reaction solution (65 μl). The LDR product (5 μl) is added to 60 μl of TMAC hybridization solutions (3 M tetramethyl-ammonium chloride, 50mM Tris-HCL [pH 8.0], 3 mM EDTA [pH 8.0], 0.1% SDS) containing 250 of each possible unique microsphere for genotyping the
Figure 2.4 Basic design of a ligation detection reaction-fluorescent microsphere assay (LDR-FMA). **A)** Essential elements of the assay include sequence (of interest)-specific LDR probes with unique 5’ TAG sequence, 5’ phosphorylated (Phos) and 3’ Biotinylated downstream probe and unique fluorescent microspheres with anti-TAG sequence that are complementary to the TAG on the sequence-specific probe. **B)** Sequence-specific and biotinylated probes were added to the PCR product of interest. Thermostable DNA ligase and buffer allowed for the probes to anneal to complementary PCR product, when present, and further hybridization occurs between the two probes. **C)** Anti-TAG sequences attached to the fluorescent microsphere hybridize with the TAG sequence extensions found at the 5’ end of the sequence-specific probe, labeling the sequence with a unique fluorescent classification microsphere. **D)** Further incubation of the TAG-LDR products with streptavidin-R-phycoerythrin (SA:PE) was performed to label the 3’ end of the sequence with SA:PE, a reporter fluorescence.
specific targets (Table 2.3). The hybridization conditions are 95°C 1.5 min (denaturation) and 37°C 25 minutes. Reporter labeling (Figure 2.4D) of the LDR product biotinylated 3’ end occurs by the addition of 6 µl of TMAC hybridization solution containing 120 ng streptavidin-R-phycoerythrin conjugate (Invitrogen, Carlsbad, CA) and incubation at 37°C for 25 minutes.

Luminex microspheres are used to classify the labeled LDR products by dual-fluorescence flow-cytometry (Bio-Plex array reader [Bio-Rad Laboratories, Hercules, CA]) and the phycoerythrin fluorescence quantifies the signal as median fluorescence intensity (MFI). This classified MFI data is collected by the Bio-Plex Manager analytical software (Bio-Rad Laboratories, Hercules, CA) and is available in an Excel (Microsoft, Redmond, WA) downloadable spreadsheet format.

G. DNA Sequence: Preparation for Analysis

Raw DNA sequence data (including nucleotide chromatograph) was obtained for each clone (Beckman Coulter Genomics, Danvers, MA) following the bidirectional sequencing. At least two clones representing each mosquito were sequenced to ensure quality and validity of the sequenced data. Two forward and two reverse reads were produced for each mosquito subjected to sequencing. These four reads were aligned (ClustalW [default settings: gap open penalty=10, gap extend penalty=6.66] (Larkin 2007)) and sequence/chromatograph quality was interrogated using Geneious 5.3 (Drummond 2009). From the ClustalW alignment a consensus sequence was generated and polymorphic sites, where peaks
representing two nucleotides were equally visible for the same site, were denoted by the use of IUPAC codes for multiple nucleotides (A or G=R, C or T=Y, A or C=M, G or T=K, C or G=S, A or T=W) (IUPAC/IUB 1974). Sequences were then assigned an identifier using an abbreviation for the species, location collected, and sample number (i.e. the first mosquito analyzed from Madang that has a species identification of An. punctulatus would receive an identifier as follows: AP_mad_01). This same identifier would be used for each subsequent gene sequenced from this mosquito in order to organize multiple gene sequences from the same organism.

Geneious was used to create further sequence alignment text files (NEXUS and FASTA) required by sequence and phylogenetic analysis programs containing all representatives for each gene. Multiple sequence alignments were performed for each gene using ClustalW in Geneious 5.3. ClustalW uses a progressive multiple alignment approach to efficiently create the most biologically relevant sequence alignment from multiple sequences. Stepwise assembly of the final sequence alignment is initiated by conducting a pairwise alignment for each combination of sequence pairs thus resulting in the creation of a distance matrix of pairwise distance scores (percent identity) computed among all sequences. This distance matrix is used to create a rough guide tree that is actually a neighbor-joining tree (discussed further in Section H). Using the relationships established in the guide tree ClustalW then sets out to realign the group of sequences by taking the two most similar sequences and creating a consensus sequence from them (preserving gapped positions). This consensus sequence is then aligned with the next closest
related sequence and a new consensus is created. This cycle is repeated until all
sequences have been included in the consensus alignment (Xiong 2006, Ye 2008).

Geneious was used to infer encoding amino acids for sequenced genes and
helpful in determining basic quantitative characteristics of the
sequences/alignments (pairwise percent identities and identical sites among
sequences, nucleotide composition). Geneious was also used to concatenate
sequence alignments from multiple genes. These concatenated alignments were
used for a specific multiple gene phylogenetic analysis, described in the following
section.

H. Basic Phylogenetic Principles and Methods

There are many objectives for using phylogenetic analysis. The purpose of
phylogenetic analysis in this dissertation is to produce a biologically meaningful
reconstruction of the evolutionary history of the Anopheles punctulatus species
complex. It is worth noting that phylogenetic analysis of multiple candidate genes
can provide an understanding of the evolution of those specific genes however the
phylogenetic relationships produced by each gene-based analysis may not
specifically represent the evolution of the species. Additionally, there are many
methods and approaches to phylogenetic analysis of genetic data.

Generally, phylogenetic analysis begins with a multiple sequence alignment
as described above. A model of substitution must then be assigned to the analysis.
This model will be used to determine how differences in the alignment are treated
or weighted. Many different models of substitution exist. For example here I will
highlight two models, one simple and one more complex. The simplest model of substitution is Jukes-Cantor, which assumes that the probability of any nucleotide substitution is equal—transitions (i.e. purine->purine; pyrimidine->pyrimidine) and transversions (i.e. purine<->pyrimidine) (e.g. the probability of replacing an A with a T [A>T] is the same as an A>G or A>C) (Xiong 2006).

A more complex model, Tamura-Nei model, assumes the probability of a transition to be higher than a transversion. Tamura-Nei also considers the frequency with which each nucleotide occurs in the sequence. For example in a sequence that has a high frequency of G, the probability of an A>G (transversion) is higher than it would have been in a sequence with equal frequency of all nucleotides. The model chosen for analysis will ultimately influence how relationships among sequences are determined (Xiong 2006, Ye 2008). A simplified alignment and relationship tree demonstrates how
these two models would treat a single point mutation in a sequence that was GC rich (Figure 2.5). Here with this alignment the probability of the A>G is greater (than it would be in a sequence with equal proportions of all nucleotides) because the G content of the sequence is high. When the Jukes-Cantor model is applied to these hypothetical sequences it may treat all three sequences equally in the second comparison. However, Tamura-Nei may establish that because the sequence has proportionally more G, a change from A>G may be more likely than an A>C change.

Deciding upon a tree-building method will then determine how relationships within the sequence alignment are visually displayed. There are numerous methods and variations of methods for tree-building and each has specific assumptions, advantages and disadvantages. There are two general categories for tree-building methods: distance-based methods and character based methods (Pevsner 2003, Xiong 2006, Ye 2008).

Distance-based tree-building methods assess evolutionary methods based upon the amount of divergence (distance) between sequences based upon a sequence alignment. A distance matrix is created and displays the distances (or differences) between each combination of sequence pairs. These methods assume that all sequences in the alignment are homologous, or similar in evolutionary origin, and that distances between two sequences can be calculated by a stepwise addition process of all the branch lengths connecting the two (referring back to the distance matrix) (Pevsner 2003, Xiong 2006).
The neighbor-joining method (NJ) is one example of a frequently used distance-based tree building method that will be employed in this dissertation. The NJ method builds a tree stepwise using the distance matrix to ascertain the two closest related sequences. The method then compares the next (third) closely related sequence to the founder clade and connects them through a node. This process is then repeated until all sequences in the alignment have been assigned a node in the tree (Pevsner 2003, Xiong 2006). NJ tree-building is useful when analyzing large numbers of sequences, as it is computationally efficient. However, since it produces a tree, and only one tree, in a stepwise manner using a distance matrix it may not capture the most accurate depiction of evolutionary relationships among the organisms.

Character-based tree-building methods do not follow an intermediate distance matrix step, but rather infer relationships between all sequences directly from characters—the sequence alignment. This inference is made by directly counting mutation characters across the entire sequence and because of this is thought to more accurately describe evolutionary relationships between sequences (Pevsner 2003, Xiong 2006).

Maximum likelihood (ML) is a commonly used example of a character-based tree-building method that will be employed in this dissertation. ML tree-building searches all possible tree topologies (at every position in the alignment) to find the most probable evolutionary path given the substitution model applied to the sequence alignment. Therefore this method is extremely computationally involved,
especially for large sequence alignments, however it may arrive at a more biologically or evolutionarily relevant tree compared to distance-based methods (Pevsner 2003, Xiong 2006).

Following the choice of tree-building methods, statistical evaluation of the inferred phylogenies display the confidence or reliability of the tree topology. Like sequence substitution methods and tree-building methods, there are several methods of tree topology evaluation. In this dissertation the bootstrapping method was used for statistical evaluation of tree topology.

Bootstrapping is a statistical method that evaluates the sampling errors that may have occurred in the process of making the phylogenetic tree and assesses the reliability of individual clades. Essentially this method repeats the tree-building process several (user defined) times, each time modifying the sequence set slightly. This method assumes a robust relationship between sequences should have enough defining characteristics that if the data set is slightly modified (for instance removing one sample, or resampling a random selection of nucleotides), that the topology if the tree should stay intact. Bootstrap values (1-100) displayed on each node of the consensus tree demonstrate the percentage of resampled trees that supported the location of that node (Felsenstein 1985, Xiong 2006). For example if a node connecting species A and species B in the bootstrap consensus tree displayed a 95 bootstrap value, and their were 1000 bootstrap replicates performed for the sequence set, then this node connecting species A and B was present in 950/1000 (95%) of the replicate trees.
Choosing a program to carry out a phylogenetic analysis can be a challenging task as there are a multitude of programs and combinations of methods that can be employed. For this dissertation, I have chosen MEGA5 software (Tamura et al. 2011) as it is user-friendly, allows for NJ (with bootstrapping) and basic ML tree-building applications and is not computationally intensive. As described previously, ML tree-building is computationally taxing and the addition of a statistical evaluation method can cause the computing time to exceed the capabilities of a personal computer. Therefore, I have chosen to utilize RAxML (Randomized Axelerated Maximum Likelihood), a program designed specifically to generate ML trees (with bootstrap values) (Stamatakis 2006, Stamatakis et al. 2008) and is freely available through online high performance computing resources like CIPRES (Cyberinfrastructure for Phylogenetic Research) (Miller 2009). RAxML also has the capability of comparing a concatenated sequence set of multiple genes where a consensus tree is formed using trees created for each gene individually. The program jModelTest v0.1.1 was utilized to aid in appropriate substitution model selection for ML analysis (Posada 2008, 2009). FigTree v1.3.1 was utilized to annotate tree files to create the figures for this dissertation (Rambaut 2009).

I. Genetic Diversity and Differentiation

Genetic diversity and differentiation was analyzed for COI sequence haplotypes. Genetic diversity of a population was assessed by calculating the heterozygosity index ($H$) using DnaSP v5 (Librado and Rozas 2009). Genetic differentiation of compared populations was assessed by calculating the fixation
index, $F_{st}$, using Arlequin v3.5 (Excoffier and Lischer 2010). Data files were formatted (i.e. haplotype information) for Arlequin analysis using DnaSP. Values for genetic diversity and genetic differentiation statistics range from 0 to 1; 0 signifying no diversity (all compared sequences have the same haplotype) and 1 signifying every sequence is of a different haplotype (Hartl and Clark 1997).
III. CHAPTER 3

Genetic Analysis of ITS2 Sequence: Evaluating Species Differences

Portions of the contents in Chapter 3 are published in the following manuscripts:


A. Introduction: Anopheles punctulatus Sibling Species Definitions

The World Health Organization’s global strategy for integrated vector management highlights the need for assessment and continuous monitoring of the vector species composition in disease transmission areas (WHO 2004). Improving the capacity to evaluate the vector species composition of a disease endemic region as well as clearly defining vectorial capacity, habitat preference, host preference etc. is contingent upon the reliability of the current species definitions. Because habitat range for mosquitoes is vast and complex, integrated vector management may require significant numbers of sampling sites and very large numbers of sampled individual mosquitoes.

As described in Chapter 1, historical cross-mating and allozyme analysis of the An. punctulatus complex demonstrated distinct differences (mating barriers, molecular polymorphisms) among members of the AP complex, suggesting the existence of multiple species, beyond the three morphotypes (An. punctulatus, An. koliensis, An. farauti s.l.) (Bryan 1973a, Bryan 1973b, Foley et al. 1993). With the development of more advanced DNA interrogation methods, species-specific differences were observed in the ITS2 and 18SrDNA sequences through analyzing the DNA sequence of one or a few representatives for each species (Beebe and Saul 1995, Beebe et al. 1999a, Beebe et al. 2000d, Benet et al. 2004b). Because of these observed genetic differences, the ‘gold standard’ method for molecular differentiation of AP sibling species was described in 1995, exploiting species-specific RFLP pattern differences in the ITS2 sequence (Beebe and Saul 1995). To
date, however, no published study has sought to test the strength and durability of the ITS2 sequence to differentiate species within the AP complex. The ribosomal RNA genes, which are inclusive of the ITS2 sequence, are localized to the X and Y-chromosomes of *Anopheles* with estimates of 470-600 copies per genome (Collins et al. 1989, Kumar and Rai 1990, Marchi and Pili 1994).

With new and on-going vector management plans in place throughout PNG (Hetzel 2009), it is important to have a firm understanding of the reliability of using ITS2 sequence to differentiate species within the AP complex. It is known that members of the AP complex display distinct but overlapping habitat preferences (Figure 1.9 (Cooper et al. 2002)). Because of these distribution patterns it has been suggested that individual species may have different preferences in hosts and biting times and that these preferences might vary between locations (Afifi et al. 1980, Bockarie et al. 1996b, Beebe et al. 2000b, Benet et al. 2004b). Since LLINs have recently been distributed to every village in the country (Hetzel 2009), monitoring species presence, behaviors, and preferences during this campaign is necessary to establish baseline distributions of species as well as correlations of species and feeding behaviors (i.e. host preference: anthropophilic vs. zoophilic, indoor [endophagic] vs. outdoor [exophagic] biting preference, indoor [endophilic] vs. outdoor [exophilic] resting preference). Given the abundance of mosquito collections that are/will be occurring with the LLIN distributions and subsequent follow-up monitoring, a robust high-throughput species identification method would allow for a more through and accurate mosquito population assessment.
Observations from Beebe and colleagues described the ITS2 sequence for one representative of each AP complex member known at the time (n=10) (Beebe et al. 1999a). Additionally, further comparison from 30 locations around coastal Australia, PNG, Solomon Islands and Vanuatu were completed for An. farauti s.s. (AFs.s.) yielding 13 additional reference sequences for AFs.s. (Beebe et al. 2000c). These reference sequences (n=23) are available in GenBank (accession numbers: AF055984, AF033213-21, AF104314-26). These two comparison studies, based on a limited sample number, suggested ITS2 sequences were highly polymorphic in a species-specific manner (Beebe et al. 1999a, Beebe et al. 2000c). Additionally, the latter study indicated polymorphism existed among members of the same species (AFs.s.) (Beebe et al. 2000c).

Given that polymorphism has been identified within a single species and that no large-scale survey of ITS2 sequences has been published for AP sibling species it is unclear how much variation exists among and between species and how reliable the ITS2 sequence is for the purpose of species differentiation. To address this problem, it was necessary to establish a diverse geographical sample set of AP complex members and to investigate the diversity of the ITS2 sequence of these representatives.

B. Methods

Details regarding sampling methods, DNA extraction, amplification and sequencing can be found in Chapter 2. For this chapter DNA sequence analysis of the ITS2 region for 90 individual mosquitoes was completed as described in Chapter
mosquitoes were used to evaluate the ITS2 LDR-FMA method, generally described in Chapter 2 and described in further detail in this chapter (Table 3.1, Figure 2.2). The ITS2 LDR-FMA was also expanded to include a probe for *Wuchereria bancrofti* (Wb) genomic DNA, detecting the presence or absence of Wb in mosquitoes. This combined assay was used to evaluate the *Anopheles* species and detect Wb in a field survey of the Dreikikir area (n=955) (Figure 2.2).

### C. Results

#### 1. ITS2 Sequence Analysis of AP Sibling Species

To establish a geographically diverse sample set of the AP species implicated in malaria and filariasis transmission in PNG, AP sibling species members were
collected from 24 villages in 7 provinces (6 mainland, 1 island) (Table 3.1). Collections were performed through activities associated with LLIN distribution (Hetzel 2009), lymphatic filariasis elimination campaigns (Thomsen et al. 2010), insecticide susceptibility trials (Keven 2010) and other ongoing studies conducted by the Entomology Unit of the PNGIMR. From the study sites sampled, 90 individual *Anopheles* that were morphologically identified as members of the AP complex, were chosen for ITS2 amplification followed by cloning and plasmid sequencing as described in Chapter 2.

Five AP complex sibling species were present in the DNA sequencing efforts: *An. punctulatus* (AP) (n=27), *An. koliensis* (AK) (n=20), *An. farauti* s.s. (AFs.s.) (n=15), *An. hinesorum* (AH) (n=12), and *An. farauti* 4 (AF4) (n=16) (*An. longirostris* [AL], a non-AP complex species, was sequenced for comparison purposes [n=4]). These five species are the target of sequence investigation in this dissertation as they are the most prevalent and widespread species either throughout the country (AP, AK AFs.s., and AH) or specifically in PNGIMR/CWRU study locations (AP, AK, AFs.s., AH and AF4), and are thought to be the main malaria vector species in PNG (Cooper et al. 2009a).

Given that only one reference sequence for each species (with the exception of AFs.s.) existed prior to this study, simple comparisons will be made between these references and the products of the current sequencing efforts for each species. In this section I will demonstrate the similarity observed among the sequenced
representatives and between reference sequences (Beebe et al. 1999a, Beebe et al. 2000c) and all representatives sequenced for this project.

All sequences were aligned in a species-specific manner using ClustalW in Geneious 5.3 as described in Chapter 2. Table 3.2 presents a summary of the sequence data. Sequences varied in size from 653 bp to 783 bp and revealed pairwise percent identity scores within each species that exceeded 98.3% (AP=98.9%, AK=98.3%, AFs.s.=99.3%, AH=99.6%, AF4=99.5%), suggesting strong sequence similarity among members of the same species.

<table>
<thead>
<tr>
<th>Species</th>
<th>n=</th>
<th>Size Range*</th>
<th>Mean Length*</th>
<th>Identity (PPI) within species</th>
<th>Identical Sites</th>
<th>Consensus vs. Reference ¹ PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>An. punctulatus</td>
<td>27</td>
<td>653-699</td>
<td>694.4</td>
<td>98.9%</td>
<td>661 (94.6%)</td>
<td>95.8%</td>
</tr>
<tr>
<td>An. koliensis</td>
<td>20</td>
<td>728-735</td>
<td>731.5</td>
<td>98.3%</td>
<td>688 (93.4%)</td>
<td>95.7%</td>
</tr>
<tr>
<td>An. farauti s.s.</td>
<td>15</td>
<td>694-701</td>
<td>698.2</td>
<td>99.3%</td>
<td>683 (97.0%)</td>
<td>97.8%</td>
</tr>
<tr>
<td>An. hinesorum</td>
<td>12</td>
<td>696-701</td>
<td>696.3</td>
<td>99.6%</td>
<td>684 (98.1%)</td>
<td>97.1%</td>
</tr>
<tr>
<td>An. farauti 4</td>
<td>16</td>
<td>768-783</td>
<td>779.9</td>
<td>99.5%</td>
<td>760 (97.1%)</td>
<td>96.4%</td>
</tr>
</tbody>
</table>

* nucleotides ¹ References from Beebe 1999b, Genbank accession nos. provided in Figure 3.1

A consensus sequence was created for each species from the multiple sequence alignment. This consensus was then aligned with the previously published reference sequence (Beebe et al. 1999a) for that species. This created five separate alignments, one for each represented species, allowing for the comparison and calculation of percent pairwise identity (PPI) between the consensus sequence generated from my dissertation and the sequence previously published by Beebe et al. (Table 3.2). These comparisons for each species revealed PPIs greater than 95.7% for all of the five species (AP=95.8%, AK 95.7%, AFs.s.=97.8%, AH=97.1%,
AF4=96.4%). These five individual species alignments were then further consolidated into one alignment to demonstrate the similarities and differences observed between the five species. This alignment allowed for a comparison of all five species, producing a PPI of 74.2% (Figure 3.1).

In this alignment of all five species the distinguishing species-specific features of the ITS2 sequence can be visualized. For example, a large insertion (31-33bp) can be found only in AF4 (at positions 487-519 bp in Figure 3.1) and an 8 bp insertion, unique to AK, is observed between 59-66 bp. Insertions like these can be observed for each species and helps to validate the ITS2 region as a sufficient DNA target for species discrimination. This alignment also helps to identify regions of homology between species, as alignments suggest that there are 353 sites where all five species share nucleotide identity.

*An. longirostris* (AL) (*Cellia*: Neomyzomyia Series), a proposed relative of the AP complex also found in PNG, was sequenced as an outgroup to facilitate comparisons for members of the AP complex. Upon inspection of the AL sequences, it is clear that little sequence homology exists between members of the AP complex and AL, as AL is significantly shorter (603 bp) than all members of the AP complex. This length difference make it is very difficult to include AL in alignments with members of the AP complex. When AL is forced into the alignment (created in Figure 3.1 by relaxing the gap penalties in ClustalW) it is observed to share
Figure 3.1 ITS2 References sequences compared with consensus sequences from the current project. Odd sequences were harvested from GenBank (accession numbers follow “#” after species identifier) and described previously (Beebe 1999a). Even numbered sequences are consensus sequences created from multiple representatives sequenced from each species. Sequences included in consensus creation were deposited in GenBank (accession nos. HM584365-HM5584454). Dashes (-) represent where sequences are in agreement with sequence #1, locations where nucleotides differ from sequence #1 are indicated by the insertion of the appropriate nucleotides and gaps in the alignment are represented as full stops (.).
significantly lower nucleotide identity (213 sites) with members of the AP complex. Pairwise percent identities comparing AL to each consensus sequence were less than 52.8% (AP vs. AL=50.9%, AK=47.5%, AFs.s.=52.3%, AH=52.8%, AF4=47.8%). This comparison with AL demonstrates further that ITS2 sequences of the AP complex members are more similar when compared to themselves than to a proposed close relative within the same taxonomical subgenus and series.

In an earlier assessment of AFs.s. diversity, Beebe and colleagues identified 13 haplotypes using RFLP patterns that were present at one or more of their 30 sampling locations. One representative from each haplotype was sequenced to illustrate that the haplotype sequence variations were geographically distributed (Beebe et al. 2000c). Aligning these 13 sequences (Figure 3.2) with the AFs.s. consensus (bottom) and original reference (top; AF055984) sequence allowed for a greater sequence comparison within the AFs.s. species. This alignment demonstrated validation of polymorphisms that had been observed between the reference and consensus in the sequence alignment of AFs.s (Figure 3.2). This comparison also identifies polymorphisms that may be isolated to a geographic region, for example an insertion (3 bp) found from 171-173 is not found in sequences from the Northern Territory of Australia (NT) and Vanuatu. This insertion is also absent in one sequence from Northern PNG, but nowhere else. When present, the insertion is polymorphic in sequence. The ‘GTG’ motif is found in two sequences restricted to Queensland and Northern PNG. ‘ACG’ is found in sequences from Rabaul, Northern PNG and South East PNG as well as the consensus sequence generated from my dissertation, where most of the AFs.s. samples were
Figure 3.2 Alignment of An. farauti s.s. representatives. Sequence #1-#14 were harvested as a sample of previously published ITS2 sequences from Genbank as described in Beebe (1999a & 2000b). GenBank Accession numbers are listed in the sequence title following “AF1”. Location of the sample collection is indicated following the accession number. Samples are from Northern Territory (NT) and Queensland (Queensl) Australia Vanuatu, and Guadalcanal (Guadal) as well as samples collected from Rabaul and along the northern (NPNG) and southeastern (SEPNG) coasts of PNG. Numbers following the location designations on these sequences refer to the specific haplotypes described in Beebe 2000b. Sequence #15 is a consensus sequence created from 15 representatives described by Henry-Halldin (2011) (Genbank accession nos. HM584365-HM584379). Dashes (-) represent where sequences are in agreement with sequence #1, differences are indicated as inserted nucleotides and gaps are represented as full stops (.).
taken from the North Coast of PNG, around Madang town. Interestingly the sequence from Guadalcanal has a 2 bp insertion (‘TG’) at this location that is not observed in any other sequence.

Figure 3.2 demonstrates that the polymorphisms observed in the sample set of AFs.s. assembled for this dissertation are shared among many members of the 13 AFs.s. haplotypes described previously. A pairwise percent identity of this alignment (98.6%) is similar to that of the pairwise percent identity for the AFs.s. sequences compiled for this project as well as intraspecific comparisons for AP, AK, AH and AF4. This bank of previously published sequences allows for the appreciation of observable polymorphisms we can expect within a species and also validates many of the polymorphisms we have observed within AFs.s. specifically.

2. Development of a High-throughput Species Diagnostic Method Utilizing the ITS2 Sequence

Observed species-specific polymorphisms, like the previously mentioned insertions (Figure 3.1), were confirmed by sequencing multiple individuals from each species, utilizing collections from different geographical regions and comparing data with previously published representative sequences. Additionally, minimal (<1.7%) polymorphism was observed among members of the same species. This suggested that certain ITS2 sequence polymorphisms might be reliable enough to exploit for species identification.

The current ‘gold standard’ molecular method for species identification is based on MspI restriction enzyme, cutting ‘CCGG’ motifs within the ITS2 sequence,
differentiating members of the AP complex by RFLP as described in Chapter 1 (Figure 1.7) (Beebe and Saul 1995). This method has been shown to be more accurate than morphological identification (Beebe and Saul 1995, Beebe and Cooper 2002, Benet et al. 2004b) because morphological characteristics can be polymorphic and cannot be used to differentiate the 8 species in the Farauti clade. The RFLP method, however, creates biohazardous waste through the use of ethidium bromide to visualize the RFLP patterns and species designations are interpreted by the researcher, a time-consuming step that introduces the possibility for human error. The RFLP method is also limited by its inability to query more than polymorphism at a time, where multiplex strategies, now available through new technology, can be utilized to simultaneously survey many regions of a gene or multiple genes.

A high-throughput method accommodating large sample sizes (in a 96-well format, automation-ready), reducing the need for gel interpretation and eliminating hazardous waste would be useful for vector population studies that are currently being conducted by PNGIMR and other mosquito control programs. To develop a high-throughput assay I utilized the general post-PCR Ligation Detection Reaction-Fluorescent Microsphere Assay (LDR-FMA) approach described in Chapter 2 (Figure 2.4), and targeted consistent species-specific polymorphisms within the ITS2 sequence for species differentiation. Species-specific probes were developed using sequence data discussed previously in this chapter. Annealing locations for classification and reporter probes are illustrated in Figure 3.3 (Henry-Halldin et al. 2011).
Figure 3.3 Internal Transcribed Spacer 2 consensus alignment for Anopheles punctulatus, An. farauti s.s., An. hinesorum, An. farauti 4 and An. koliensis demonstrating locations of LDR-FMA species identification probes. Consensus sequence for each member was created using ITS2 sequence from each species was described previously. Dashes (-) represent where sequences are in agreement with 'APConsens', differences are indicated as inserted nucleotides and gaps are represented as full stops (.) Annealing positions of species classification probes which include TAG sequences specific to FlexMap classification microspheres are indicated in grey. Annealing positions of reporter probes with a biotin reporter extension are immediately upstream from classification probe locations and are indicated in black. This figure was reproduced with permission from American Journal of Tropical Medicine and Hygiene, and appears in Henry-Halldin et al. 2011.
To assess sensitivity and specificity of the LDR-FMA probes, DNA sequence controls were established for each species using extracted plasmid DNA from sequenced clones described previously in this chapter. It is worth mentioning that female mosquitoes collected from the field may be infected with parasites and also may have taken a bloodmeal (Burkot et al. 1988, Burkot et al. 1990, Cooper et al. 2009a). To ensure that the LDR probes did not cross-hybridize to non-mosquito DNA that may be present in the whole mosquito DNA extraction, a DNA blood sample along with DNA from the four human malaria Plasmodium species and Wuchereria bancrofti were included in the evaluation of LDR probe specificity (Figure 3.4). The AP sibling species probes detected each of the five species in a

![Figure 3.4 Detection of species-specific DNAs by the Anopheles species multiplex PCR-LDR-FMA. Data represents a summary of twelve individual control experiments detecting AP, AK, AFs.s. (AF), AH and AF4 as well as Plasmodium falciparum (Pf), P. vivax (Pv), P. malariae (Pm), P. ovale (Po), Wuchereria bancrofti (Wb) and human (HBS) genomic DNAs. Whereas genomic DNAs were added individually into the control reaction, LDR and FMA reactions included oligonucleotide probes representing all five Anopheles species. Numbers in parentheses next to species designations in the legend ((C1), (11), (59), (68), (78)) identify Luminex FlexMap™ microspheres (Luminex Corp., Austin, TX). 'No DNAs' identifies a blank sample to which no genomic DNAs were added. This figure was reproduced with permission from American Journal of Tropical Medicine and Hygiene, and appears in Henry-Halldin et al. 2011.](image-url)
species-specific manner with an average mean fluorescence intensity (MFI) >5,000 (AP > 5,000, AK > 11,000, AFs.s. > 20,000, AH > 21,000, AF4 > 5,000) and background MFI below 900. Additionally, MFI signals from the parasite and human DNAs were well below 900 MFI. This confirmed that the presence of these non-mosquito DNAs in a mosquito DNA extraction would not impede mosquito species identification.

3. Application of ITS2 LDR-FMA

a. Evaluation of species identification with established methods

The ITS2 LDR-FMA was tested on collection of mosquitoes (n=340) from the sample locations described previously (Table 3.1) to assess their capacity for differentiating species in geographically diverse populations. Morphological identification of the 340 mosquitoes had been completed prior to DNA extraction as described in Chapter 2. Concordance between the ITS2 LDR-FMA and morphological species identification is demonstrated in Table 3.3. Because members of the An. farauti clade are morphologically indistinguishable, morphology cannot classify members of the AP complex beyond three categories: An. punctulatus, An. koliensis and An. farauti s.l. The ITS2 LDR-FMA demonstrates the capacity to identify five species of the AP complex—An. punctulatus, An. koliensis, and three members of the Farauti clade: An. farauti s.s., An. hinesorum, An. farauti 4.

Of the 67 mosquitoes morphologically identified as AP, 10 were reclassified as AK by the ITS2 LDR-FMA (85% concordance between morphology and ITS2 LDR-FMA). Of the 115 mosquitoes morphologically identified as AK, ITS2 LDR-FMA
reclassified 34
(AP=11, AFs.s.=15, 
AH=3 and AF4=5)
(70% concordance).

Of the 158 morphologically identified as a member of the Farauti clade (AF s.l.), ITS2 LDR-FMA reclassified 37 (AP=33, AK=4) and the remaining 121 were further differentiated as AFs.s. (n=26), AH (n=49) or AF4 (n=46) (76.5% concordance). Overall concordance between morphology and ITS2 LDR-FMA was 76.2%.

A subset of the 340 mosquitoes (n=117) was subjected to the ITS2 RFLP species diagnostic by the Entomology Unit of the PNGIMR without prior knowledge of ITS2 LDR-FMA results. This data was compared to ITS2 LDR-FMA results and demonstrated a 98.3% concordance (Table 3.4). Comparison between ITS2 RFLP and morphology demonstrated similar concordance (76%) as ITS2 LDR-FMA vs. morphology.

Table 3.3
Concordance assessment between morphology* and ITS2 classification† for An. punctulatus sibling species

<table>
<thead>
<tr>
<th>LDR-FMA Species ID</th>
<th>Morph. ID n=</th>
<th>AP</th>
<th>AK</th>
<th>AF s.s.</th>
<th>AH</th>
<th>AF4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>67</td>
<td>57</td>
<td>10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>115</td>
<td>11</td>
<td>81</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>AF s.l.</td>
<td></td>
<td>158</td>
<td>33</td>
<td>4</td>
<td>26</td>
<td>49</td>
</tr>
</tbody>
</table>

*AFs.l.=An.farauti s.l. (specific species morphologically unidentifiable) †AP=An. punctulatus s.s., AK=An. koliensis, AFs.s.=An. farauti s.s., AH=An.hinesorum, and AF4=An. farauti 4. Reproduced with permission from the American Journal of Tropical Medicine and Hygiene and appears in Henry-Haldin et al. 2011.

Additionally, ITS2 DNA sequence was also obtained for 13 of the 117 mosquitoes subjected to both ITS2
RFLP and ITS2 LDR-FMA. One hundred percent concordance between ITS2 LDR-FMA, RFLP and reference sequence comparison was observed for these 13 mosquitoes (AP=5, AK=6, AFs.s.=2) (Henry-Halldin et al. 2011).

b. Multiplex evaluation of mosquito species and infection with *W. bancrofti* detection

Given the concordance of the ITS2 LDR-FMA to the ‘gold standard’ molecular species identification method and the demonstrated ability to distinguish members of the AP complex from different geographic regions of PNG, the newly designed method was applied to a field collection set of individual mosquito DNA extractions from the Dreikikir region (n=1056) (Table 3.5), an area that *W. bancrofti* (Wb) filariasis is endemic (Thomsen et al. 2010). This collection had previously been subjected to morphological identification and a portion had ITS2 RFLP species identification completed. For the interrogation of this sample set the ITS2 LDR-FMA method was expanded to a multiplex assay including detection probes designed to capture Wb DNA (Mehlotra et al. 2010). The multiplex reactions were completed while blinded from RFLP or morphological species identification data. The ITS2 DNA fragment amplified for 955 samples. Each of these was identified as one of the five AP species using ITS2 LDR-FMA. This sample set demonstrated an 80.2% concordance with
morphological identification and 95.6% concordance with ITS2 RFLP. Wb DNA was detected in 19.1% of the mosquitoes in the survey. This exercise demonstrated the reproducibility of concordance between ITS2 RFLP and LDR-FMA and the application of the ITS2 LDR-FMA in a field epidemiology study. This exercise also highlighted the versatility and forward expansion of the LDR-FMA format allowing the ITS2 species diagnostic to be multiplexed with Wb detection.

D. Discussion

For members of the An. punctulatus species complex only one ITS2 reference sequence was available in GenBank (with the exception of An. farauti s.s.). Since very little ITS2 sequence data has been published for the AP complex, it was unclear, prior to the current project, how reliable the ITS2 sequence was for species differentiation. Additionally, because intraspecies variability has only been described for An. farauti s.s. (Beebe et al. 2000c), the degree of polymorphism among most of the AP complex has not been established. Because ITS2 has become the common reference sequence for molecular species identification, it became necessary to characterize consistency and diversity of this sequence among AP sibling species from geographically diverse regions in Papua New Guinea.

This chapter describes characteristics of 90 newly sequenced ITS2 representatives of the AP complex from a various locations within Papua New Guinea (Figure 2.1 & Table 3.1). The ITS2 region of each representative was sequenced as described in Chapter 2 and compared among themselves and with previously published references (Beebe et al. 1999a, Beebe et al. 2000c). Comparing
sequences obtained from each species demonstrated >98% similarity within each species (Table 3.2 Pairwise % Identity). Each group of species had identical sequence at least 93.4% of nucleotide sites (Table 3.2 Identical Sites).

A consensus sequences created for each species enabled comparison with previously published reference sequences (Table 3.2), interspecies comparisons and visualization of species-specific polymorphisms (Figures 3.1 and 3.2). Greater than 95.7% similarity existed between the published references sequences available and the consensus for each species (Table 3.2). Figure 3.1 provides two different levels of comparison. First it demonstrates the similarity of the reference and consensus sequence for each species. Secondly it displays regions within the ITS2 sequence that are shared among all five species and those that are unique to a specific species.

Figure 3.2 highlights the intraspecies polymorphisms that have been described for AFs.s., demonstrating a 98.6% pairwise identity for previously analyzed AFs.s. representatives. These earlier results are consistent to the percent pairwise percent identity (99.3%) observed for the 15 newly sequenced AFs.s. representatives from the current project. This variation among species (Table 3.2) is not surprising. The rDNA, within which the ITS2 sequence occurs, is thought to be repeated 470-600 times within the genome (Collins et al. 1989, Kumar and Rai 1990, Marchi and Pili 1994) and even though it is thought that concerted evolution acts to maintain sequence homology within individuals, slight variation within populations of species has been shown to occur (Dover et al. 1982, Liao 1999, Ganley and Kobayashi 2007).
Given that the observed species-specific polymorphism from Figures 3.1 and 3.2 appear to be reliable for species discrimination between species an effort was made to develop a high-throughput molecular species identification method. Species-specific DNA probes were designed to target these polymorphic regions for this new method. These species-specific probes demonstrated high sensitivity and specificity for hybridization to the intended species (Figure 3.4) and little background for unintended species; unintended hybridization to non-mosquito genomic DNAs (human, \textit{Plasmodium, W. bancrofti}), that might be found within the mosquito, was not observed.

This ITS2 LDR-FMA method was applied to a geographically diverse collection (n=340) and was shown to be 76.2\% concordant with morphological identification (Table 3.3). Morphological identification can only distinguish members of the AP complex as one of three species and has been shown to be between 50-75\% accurate when compared to molecular diagnostic methods in previous studies (Foley et al. 1993, Cooper et al. 2002, Benet et al. 2004b). Additionally, certain members of the AP complex have been shown to display highly polymorphic coloration patterns of the proboscis (AF4 and AK) and are frequently misclassified especially in areas that multiple species habitats are overlapping (Cooper et al. 2002, Benet et al. 2004b). Therefore the resulting concordance between morphological identification and ITS2 LDR-FMA is not surprising.

When ITS2 LDR-FMA was compared with the existing ‘gold standard’ molecular identification method (ITS2 RFLP) it was found to be 98.3\% concordant
This concordance provided confidence that the ITS2 LDR-FMA was an efficient, accurate, high-throughput method of identifying species for large collections of mosquitoes. These results provided further incentive to incorporate Wb-detection probes into the multiplex assay to facilitate an assessment of Wb infection in a large field survey collection of mosquitoes from a filariasis-endemic region of Papua New Guinea.

Applying this method to a field collection of 1056 samples demonstrated its usefulness and applicability to an entomological field evaluation of *Anopheles* species composition and Wb infection. These application exercises also demonstrated that the polymorphisms, which had been targeted for species-specific DNA probe hybridization, were consistently distributed across sampled locations within PNG. Many of the sampled locations are several hundred kilometers apart and the same species inhabit many of these isolated sites. As we can assume that since the average mosquito may only travel 1-2 km over a lifetime (Le Prince et al. 1917, Charlwood et al. 1988) there may be reduced gene flow between widely separated and potentially isolated populations. This suggests that these targeted polymorphisms are possibly ancient and may have been fixed within the individual species prior to their isolation.

The ITS2 LDR-FMA allows for the interrogation of the identified species-specific polymorphisms in large sample sets. If, in the course of a vector population study, a tested individual would lack a MFI signal or the signal would be abnormally low for species identification, this sample could be sequenced to investigate the ITS2
sequence composition. Unexpected LDR-FMA signal could result from undescribed polymorphism or may be an AP complex member for which a reliable DNA probe has not been developed (i.e. An. torresiensis, An. farauti 5, 6 and 8, An. irenicus, and An. species near punctulatus). Therefore this method allows for expansion to interrogate other polymorphisms within ITS2 as well as the addition of probes to detect the remaining members of the AP complex.

Finally, it is common knowledge that detecting Wb in the human requires nighttime blood draws as Wb circulates in the bloodstream at night only. Furthermore, obtaining human blood samples can be very time consuming and requires informed consent and skilled phlebotomists. Collecting large numbers of mosquitoes, on the other hand, involves much less administrative permission, can be carried out at any time of day and possibly requires less manpower and resources. Therefore detecting Wb in the mosquito population would serve as the optimal tool for monitoring Wb in a human population. Xenomonitoring for the parasite among vectors can also be employed as an end-stage monitoring strategy for disease elimination programs—an essential component to ensure sustained elimination. The versatility and expandability of the ITS2 LDR-FMA method was demonstrated simultaneous identification of Anopheles species and presence of Wb infection in a field survey from a filariasis endemic region of PNG. Application of this multiplex assay will greatly aid in filariasis elimination studies as well as monitoring efforts following vector and parasite management interventions, such as LLINs or mass anti-filarial drug administrations.
As discussed in Chapter 1, vector species compositions and habits are highly variable across disease-endemic regions. This method will greatly aid integrated vector management projects by increasing the entomological capacity of studies involving vector species composition and ecology in disease endemic regions. Having the ability to assign species identification and detect the presence of parasites in large collections of mosquitoes will allow entomologists to better understand specific species habits and habitat preferences, improving species-specific knowledge of vectorial capacity, biting times and preferences, breeding site and oviposition preferences, mosquito species-parasite associations etc. Ultimately, once a clear understanding of these vector characteristics are achieved, a more successful and sustained vector management strategy will contribute positively to human disease elimination in a targeted location.
IV. CHAPTER 4

Interrogation of the Voltage Gated Sodium Channel gene:
Evaluating Species Differences and Monitoring Mutations Associated with Insecticide Resistance

Portions of the contents in Chapter 3 are published in the following manuscript:

**A. Introduction: VGSC, Insecticide Resistance Mutations and the *An. punctulatus* Complex**

In response to insecticide pressure, many insects from aphids to cockroaches and including *Anopheles*, have developed resistance to the chemicals used for insecticidal purposes (Martinez-Torres D 1998, Martinez-Torres et al. 1999, Hemingway and Ranson 2000, Ranson et al. 2000, Soderlund 2008). Through intensive insecticide exposure for vector control purposes, malaria vectors in many regions of the world have developed resistance to a number of different chemical compounds, most commonly DDT and pyrethroids. Resistance phenotypes have included a variety of mechanisms such as enhanced metabolic detoxification of insecticides, as well as behavioral adaptation avoiding contact with insecticides (Brogdon and Barber 1990, Prapanthadara and Ketterman 1993, Roberts and Andre 1994, Vulule et al. 1999, Ranson and Hemingway 2004). Insecticide resistance has also been associated with genetic variation, including variable patterns of gene expression and single nucleotide polymorphisms (SNPs) (Martinez-Torres D 1998, Weill et al. 2002, Du et al. 2005, Syafruddin et al. 2010). Recent studies focusing on the over-expression of detoxification genes in mosquitoes have demonstrated that resistant populations can harbor multiple resistance mechanisms and this observed over-expression can be coupled with resistance-associated SNPs (Brooke 2008, Djouaka et al. 2008, Awolola et al. 2009, Brooke and Koekemoer 2010).

Because several different insect species have shown evidence for development of specific polymorphisms in the Voltage Gated Sodium Channel gene
(VGSC) when exposed to DDT and pyrethroid insecticides (Williamson et al. 1996, Vais et al. 2001, Foster et al. 2003), this gene has become a starting point to investigate insecticide pressure and development of resistance (Martinez-Torres D 1998, Hemingway and Ranson 2000, Soderlund 2008). The VGSC gene is located on chromosome 2L in Anopheles. In mosquitoes, knockdown resistance (kdr), associated with exposure to DDT and pyrethroids that target the VGSC, has been associated with an A>T mutation (TTA>TTT; L1014F) (Martinez-Torres D 1998) or a T>C mutation (TTA>TCA; L1014S) (Ranson et al. 2000) in codon 1014 of VGSC (Figure 4.1). These SNPs are thought to cause an inactivation of the neuronal VGSC by inducing conformational changes, reducing the affinity of DDT and pyrethroid to bind with the VGSC. Reduced insecticide binding to the VGSC enables the insect to withstand exposure and avoid paralysis and death (Chang and Plapp 1983, Vais et al. 2001, O’Reilly et al. 2006, Davies et al. 2007).

The L1014F SNP has been observed in African, Indian and Asian malaria vectors (Reimer et al. 2008, Santolamazza et al. 2008, Nwane et al. 2009, Singh et al. 2010, Verhaeghen et al. 2010), and most recently in vector Anopheles of neighboring Indonesia (Syafuddin et al. 2010). The L1014S SNP has only been described for An. gambiae in Africa and An. culicifacies in India (Ranson et al. 2000, Reimer et al. 2008, Singh et al. 2010). Furthermore, genetic evaluation of the VGSC in An. gambiae populations harboring the kdr mutations has demonstrated that the L1014F mutation has evolved independently multiple times within this species. Additional data suggests that L1014F has introgressed (moved from one species/subspecies to another by hybridization and repeated backcrossing) from the An. gambiae M form
into the S form on at least one occasion (Etang et al. 2009). These combined observations suggest that following insecticide pressure, the kdr mutation has developed independently in different species across the world in a relatively rapid manner. Many other point mutations in the coding region surrounding kdr have been described and are associated with very high levels of resistance. For example the super-kdr mutation (M918T) (Figure 4.1) confers increased pyrethroid

---

**Figure 4.1 Arthropod Voltage Gated Sodium Channel and region of interest.\(^A\)** VGSC 4 domain transmembrane protein channel. Domain II is the focus of this dissertation. \(^B\) Domain II S4-S6 are targeted for PCR amplification (green arrows are forward and reverse primers) as this region contains the M918T super-kdr mutation site (white circled X) and L1014F/S kdr mutation site (red X). \(^C\) VGSC PCR product description for An. punctualts species complex containing 3 exons coding for portions of S4-S6 separated by 2 variable sized introns, ‘Intron 1’ and ‘Intron 2’. Locations of super-kdr and kdr mutation sites are identified.
resistance and has been described in many insects, including *Culex*. To date, however these mutations have not yet been detected in *Anopheles* (Williamson et al. 1996, Guerrero et al. 1997, Vais et al. 2001, Davies et al. 2007).

As described in Chapter 1, PNG has a long, variable and site-specific history of insecticide use for malaria vector control purposes. Documented vector control campaigns started with DDT use during World War II. Indoor residual spraying (IRS) of DDT and dieldrin persisted throughout the WHO malaria eradication campaign of the 1950s and 60s (reportedly reaching 50% coverage) and continued through the 1980s invariably and the introduction of ITNs and LLINs throughout the late 1980s, 90s and 2000s has eventually reached the entire country over the last five years (Peters 1960c, a, Avery 1974, Gunther 1974, Parkinson 1974, Genton et al. 1994, Bockarie et al. 2002, Hetzel 2009).

However, given this long and, at times, intense usage of insecticides in PNG, very little has been done to evaluate the insecticide susceptibility of the vector population. A recent survey of AP complex members in five PNG locations has demonstrated complete susceptibility, according to WHO standards, to LLINs (deltamethrin-treated) and lambda-cyhalothrin (used for IRS purposes). The recent survey was conducted on 883 female adult mosquitoes collected as larvae in the wild (AP=652, AFs.s.=206, AH=25) (Keven et al. 2010). This study is the first report in recent years of the AP complex insecticide susceptibility in PNG and demonstrated all three species in the surveyed locations were susceptible to both insecticides according to WHO standards. Previously, *An. punctulatus, An. farauti* s.l.
and *An. koliensis* in the ‘Papuan’ region were shown to be susceptible to DDT concentrations of 0.5% or lower, however it is unclear how these species were evaluated or where in the South East Pacific these observations originated as detailed published data are not available (Chow 1970).

To date, Southeast Pacific malaria vectors have not been surveyed for mutations associated with insecticide resistance. Since equal exposure to insecticides is unlikely, development, prevalence, and distribution of these mutations may occur independently among members of the AP complex. Therefore performing species-specific interrogation of the *VGSC* DNA sequence would provide surveillance of mutations most commonly associated with resistance to DDT and pyrethroids. Evaluation of the *VGSC* sequence would also provide insight into the *VGSC* intronic sequence (Figure 4.1), which has proved helpful for answering gene flow and evolution questions related to other *Anopheles* mosquito species (Etang et al. 2009). Since the mosquito samples used in this dissertation were all collected prior to the recent countrywide distribution of LLINs (Hetzel 2009), this sequence information will provide important baseline data on *VSGC* polymorphism preceding significant selective pressure of insecticides.

Additionally, I have been given access to a subset of genomic DNAs from the mosquitoes evaluated by Keven and colleagues for insecticide susceptibility, described above (Keven et al. 2010). These samples will provide a useful comparison between the insecticide susceptibility phenotype and *VGSC* genotype. Finally, sequencing *VGSC* of multiple AP complex members will provide both coding
and non-coding nuclear sequence that can be used for comparison of species-specific DNA polymorphism with previously described nuclear sequence (ITS2) in this dissertation.

**B. Methods**

Details regarding sampling methods, DNA extraction, amplification and sequencing can be found in Chapter 2. For this chapter DNA sequence analysis of the \textit{VGSC} for 89 individual mosquitoes was completed as described in Chapter 2 (Table 4.1, Figure 2.2). Additionally 312 mosquitoes were used to evaluate the \textit{VGSC} LDR-FMA method, generally described in Chapter 2 and described in further detail in this chapter (Table 4.1, Figure 2.2).

A site-directed mutagenesis method was utilized to introduce a point mutation in a \textit{VGSC} PCR representative creating a positive genomic control possessing the kdr mutation (TTT) (Stratagene QuikChange® Site-Directed Mutagenesis Kit, Wilmington, DE) in accordance with the manufacture’s protocol. These samples were then sequenced confirming the introduction of the kdr point mutation (TTT) was successful.

**C. Results**

1. **Sequencing and analysis of VGSC for \textit{An. punctulatus} sibling species**

DNA sequence analysis was performed for multiple alleles for the five AP sibling species that are the focus of this dissertation: AP (n=31), AFs.s. (n=14), AH (n=13), AF4 (n=12) and AK (n=20) (Table 4.1). These ninety individual mosquitoes
were subjected to PCR amplification and DNA sequence analysis of a portion of the VGSC using a variation of previously published methods (Martinez-Torres D 1998, Weill et al. 2000) as described in Chapter 2. This method captured a ~1350 bp region of the VGSC where mutations associated with knockdown (kdr) and super knockdown resistance (super-kdr) have been identified among many arthropods (Figure 4.1).

This 1350 bp region includes sequence encoding the 4th to 6th transmembrane segments (S4-S6) of VGSC domain II; codons 908 to 1026 (Williamson et al. 1996, Martinez-Torres D 1998) including two introns, referred to as Intron 1 and Intron 2. An. longirostris representatives (Table 4.1), co- endemic to PNG, were collected and sequenced for use as a comparison group since no other full-length sequences were available, either within the AP complex or other Anopheles mosquitoes, in Genbank for comparison.

<table>
<thead>
<tr>
<th>Site</th>
<th>Province</th>
<th>Species Evaluated*</th>
<th>P</th>
<th>F1</th>
<th>H</th>
<th>F4</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albulum</td>
<td>EAST SEPIK</td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dreikir</td>
<td>EAST SEPIK</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nale</td>
<td>EAST SEPIK</td>
<td></td>
<td>11</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nanaha</td>
<td>EAST SEPIK</td>
<td></td>
<td>14</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nghambule</td>
<td>EAST SEPIK</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peneng</td>
<td>EAST SEPIK</td>
<td></td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yawatonga</td>
<td>EAST SEPIK</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilbil</td>
<td>MADANG</td>
<td></td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimer</td>
<td>MADANG</td>
<td></td>
<td>12</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hudini</td>
<td>MADANG</td>
<td></td>
<td>8</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maraga</td>
<td>MADANG</td>
<td></td>
<td>10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naru</td>
<td>MADANG</td>
<td></td>
<td>12</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sausi</td>
<td>MADANG</td>
<td></td>
<td>11</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yagaua</td>
<td>MADANG</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lorengau</td>
<td>MANUS</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gingala</td>
<td>MOROBE</td>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Godowa</td>
<td>MOROBE</td>
<td></td>
<td>2</td>
<td>21</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muneng</td>
<td>MOROBE</td>
<td></td>
<td>36</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ramu</td>
<td>MOROBE</td>
<td></td>
<td>1</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitapena</td>
<td>WEST SEPIK</td>
<td></td>
<td>6</td>
<td>1</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yapsie</td>
<td>WEST SEPIK</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kuru</td>
<td>WESTERN</td>
<td></td>
<td>1</td>
<td>1</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brimba</td>
<td>WHP**</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singoropa</td>
<td>WHP**</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Number of individual mosquitoes included in this study; P=An. punctulatus, F1=An. farauti s.s., H=An. hinesorum, F4=An. farauti 4, K=An. koliensis. †Villages represented by VGSC’s sequence data §Locations of insecticide susceptibility trials, ¶An. longirostris collected in these locations for genetic comparison purposes, **WHP= Western Highlands Province,
a. General Sequence Observations

Alignment of 31 alleles of *An. punctulatus* (pairwise percent identity 98.5%; Genbank Accession no. HQ173943-HQ173973), 14 of *An. farauti s.s.* (99.4%; HQ173904-HQ173917), 13 of *An. hinesorum* (98.1%; HQ173918-HQ173930), 12 of *An. farauti 4* (99.0%; HQ173931-HQ173942), and 20 of *An. koliensis* (98.4%; HQ173974-HQ173993). Among all 90 alleles the kdr (L1014F/S) and super-kdr (M918T) mutation sites were identified. There was no evidence of the kdr (L1014F/S) or super-kdr (M918T) mutations.

Consensus sequences were created for each of the five AP complex sibling species using ClustalW in Geneious 5.3. General comparisons within species and between consensus sequences and *An. longirostris* (AL) (HQ173903) are described in Table 4.2, where PPI (pairwise percent identity) within species is greater than 98.1% and PPI between consensus representatives and AL ranged between 78.3% (AK consensus vs. AL) an 80.4% (AH consensus vs. AL). PPI among the consensus sequences of the five AP complex members was 89.8%.

<table>
<thead>
<tr>
<th>Species</th>
<th>n=</th>
<th>Size Range*</th>
<th>Mean Length*</th>
<th>Pairwise Identity (PPI)</th>
<th>Identical Sites</th>
<th>Consensus 1(^1) vs. AL 2(^2) PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>An. punctulatus</em></td>
<td>31</td>
<td>1303-1346</td>
<td>1333.7</td>
<td>98.5%</td>
<td>1186 (87.9%)</td>
<td>79.3%</td>
</tr>
<tr>
<td><em>An. farauti s.s.</em></td>
<td>14</td>
<td>1351</td>
<td>1351.0</td>
<td>99.4%</td>
<td>1311 (97.0%)</td>
<td>79.7%</td>
</tr>
<tr>
<td><em>An. hinesorum</em></td>
<td>13</td>
<td>1366-1377</td>
<td>1372.0</td>
<td>98.1%</td>
<td>1274 (92.5%)</td>
<td>80.4%</td>
</tr>
<tr>
<td><em>An. farauti 4</em></td>
<td>12</td>
<td>1349-1362</td>
<td>1351.7</td>
<td>99.0%</td>
<td>1309 (95.9%)</td>
<td>79.3%</td>
</tr>
<tr>
<td><em>An. koliensis</em></td>
<td>20</td>
<td>1366-1379</td>
<td>1375.2</td>
<td>98.4%</td>
<td>1279 (92.5%)</td>
<td>78.3%</td>
</tr>
</tbody>
</table>

* indicates nucleotides, \(^1\) Consensus sequence for each species, \(^2\) *An. longirostris* (AL) reference sequence (HQ173903)
b. Coding region sequence comparisons

i. General comparisons

Because sequence conservation is greater in the exon than introns *VGSC* sequence comparisons were made specifically among coding regions of the 90 sequenced AP complex members to define variable sites within the sequenced coding region. These comparisons were made between AP complex members and fragments from *An. dirus* (AD) (236 bp, partial exon; Genbank Accession no. DQ026439) from Southeast Asia/Malaysia, and *An. gambiae* (AG) (407 bp, partial exon; Y13592) from Africa.

| *An. punctulatus* species complex* | *VGSC* coding sequence comparisons | Overall, among species comparisons revealed high similarity with PPI >99.1% (Table 4.3). Comparing all members of the AP complex revealed that all members shared identical nucleotides at 305/358 sites (85.2%) and display PPI values of at least 98.4%.

<table>
<thead>
<tr>
<th># bp of overlap</th>
<th>Range**</th>
<th>Identical Sites (%)</th>
<th>Pairwise % Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP (n=31)</td>
<td>358</td>
<td>1-358</td>
<td>340 (95.0%)</td>
</tr>
<tr>
<td>AFs.s. (n=14)</td>
<td>358</td>
<td>1-358</td>
<td>346 (96.6%)</td>
</tr>
<tr>
<td>AH (n=13)</td>
<td>358</td>
<td>1-358</td>
<td>346 (96.6%)</td>
</tr>
<tr>
<td>AF4 (n=12)</td>
<td>358</td>
<td>1-358</td>
<td>345 (96.4%)</td>
</tr>
<tr>
<td>AK (n=20)</td>
<td>358</td>
<td>1-358</td>
<td>343 (95.8%)</td>
</tr>
<tr>
<td>AP Complex (n=90)</td>
<td>358</td>
<td>1-358</td>
<td>305 (85.2%)</td>
</tr>
<tr>
<td>AP Complex v. AL†</td>
<td>358</td>
<td>1-358</td>
<td>301 (84.1%)</td>
</tr>
<tr>
<td>AP Complex v. AD†</td>
<td>213</td>
<td>146-358</td>
<td>171 (80.3%)</td>
</tr>
<tr>
<td>AP Complex v. AG†</td>
<td>349</td>
<td>10-358</td>
<td>281 (80.3%)</td>
</tr>
<tr>
<td>AL v. AD</td>
<td>213</td>
<td>146-358</td>
<td>200 (93.9%)</td>
</tr>
<tr>
<td>AL v. AG</td>
<td>349</td>
<td>10-358</td>
<td>322 (92.3%)</td>
</tr>
<tr>
<td>AD v. AG</td>
<td>213</td>
<td>146-358</td>
<td>186 (87.3%)</td>
</tr>
</tbody>
</table>

---

*AP= *An. punctulatus*, AFs.s.= *An. farauti* s.s., AH= *An. hinesorum*, AF4= *An. farauti* 4, AK= *An. kolensis*. § Base pairs. **bp using AP coding region as reference; † AL= *An. longirostris* (HQ173903), AD= *An. dirus* (DQ026439), AG= *An. gambiae* (Y13592)
with all sequences sharing identical nucleotides at 301/358 sites (84.1%) and a PPI of 98.4%.

Available reference sequences for *An. gambiae* (AG) and *An. dirus* (AD) were aligned with the 90 AP complex representatives. The *An. gambiae* and *An. dirus* sequences, however, did not span the entire 358 bp coding region that was sequenced for the AP sibling species. Therefore comparisons with these sequences are described in Table 4.3 indicating the number of overlapping base pairs considered in the comparisons. For AG 349 bp, spanning the region from nucleotide 10-358 (using the AP complex consensus as the reference) was available for comparison. AG and the AP complex shared 281/349 identical sites (80.5%) and have a PPI of 98.4%. The available reference sequence for AD of 213 bp spanning nucleotides 146-358, was considerably shorter with respect to the region of overlap with AP complex. AD comparison with AP complex revealed 171/213 identical sites (80.3%) and a PPI of 97.8%. Additional comparisons were performed among the outgroup AL, AG and AD sequences. The comparison between AD and AG produced the lowest similarity among the three comparisons (186/213 identical sites [87.3%], PPI 87.3%). When comparing the AP complex to AL, AD and AG at once (213 sites) only 160 sites are shared (75.1%) with a PPI of 97.5%.

Not surprisingly results for the *VGSC* coding region examined show that the AP consensus shares the greatest sequence homology with AL, followed by AG and AD. Comparisons between AP complex and AG and AD, respectively, produce similar percentages of identical sites and pairwise identity. It is difficult to ascertain
the significance of this observation as these comparisons are being made using partial sequences and only a single representative for AG and AD.

ii. **SNP identification**

Examination of the VGSC coding region (358 bp) among the AP complex representatives revealed fifty-three variable sites where fifty-six unique mutations were found. Among these, twenty SNPs were observed repeatedly within and between the AP complex sibling species; thirty-six SNPs were observed only once. Given potential for taq-polymerase errors we performed a separate evaluation of mutations observed in only one sequence. While the calculated error rate \(1.08 \times 10^{-3}; 36/32,310\) nucleotides) is higher than standard amplitaq error rates \([2.0 \times 10^{-5}\) to \(2.1 \times 10^{-4}\) /bp (Keohavong and Thilly 1989, Lundberg et al. 1991)], it was not possible to validate these thirty-six SNPs through repeated observation in this sequence survey. Therefore they were not considered in subsequent analyses.

Sixteen of the twenty repeated SNPs were synonymous mutations (all occupying the third position of the codon) (Table 4.4), ten of these sites (codons 909, 910, 924, 956, 973, 987, 996, 999, 1007, 1026) occurred in multiple AP sibling species. Codons 909 (G>C) and 910 (G>A) were found to harbor mutations that occurred together and separately in each species with the exception of AFs.s.s. where the mutations were not observed together. The predominant 909-910 haplotype was ‘GG’ occurring in 42/90 sequenced alleles (AP, n=17; AK, n=12; AF4, n=6). In AFs.s.s. CG was the predominant 909-910 haplotype (n=7) and for AH, the ‘GA’ and ‘CG’ SNP combinations were observed most often and equally (n=4 for each).
codon 1026 a C>A SNP was observed in 30/90 sequenced alleles (AP=10, AK=8, AF s.s.=5, AH=5, AF4=2). Six of these repeatedly observed synonymous mutations were shared between two species. At codon 924 a G>A SNP was observed once in both AFs.s. and AH; at 956 a C>T transition was observed in AH (n=2) and AP (n=2); at 973 a T>C transition was observed in AFs.s. (n=1) and AK (n=1); at 987 a T>C transition was observed at fixation in both AF4 (n=12) and AP (n=31); at 996 C>T was observed at fixation in both AF4 (n=12) and AP (n=31); 999 C>T was observed in AF4 (n=1) and AK (n=20), and A>C at 1007 was observed in AFs.s. (n=11) and AP (n=30). The remaining 6 repeatedly observed synonymous mutations observed in 5 codons (codons 962, 964, 980, 981, and 985) were species-specific. At codon 962 a C>T transition was observed in two AF4 alleles; at 964 an A>G transition was observed at fixation in AK (n=20); at 980 a T>G transversion was observed at fixation in AF4 (n=12) while a T>C transition was observed at fixation in AK (n=20); at 981 an A>G transition was observed at fixation in AF4 (n=12), and at 985 an A>G

Table 4.4

<table>
<thead>
<tr>
<th>Codon</th>
<th>909, 910</th>
<th>924</th>
<th>946</th>
<th>962</th>
<th>964$</th>
<th>973</th>
<th>980</th>
<th>985</th>
<th>995</th>
<th>996</th>
<th>999</th>
<th>1001</th>
<th>1007</th>
<th>1026</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/N 1</td>
<td>S S S S N S S N S S S S</td>
<td>S S S S N S S S S N S S S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP (n=31)</td>
<td>5 7 2 - 2 1 - - - 1 - - - 31 31 1 31 - - 30 1 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFs.s. (n=14)</td>
<td>7 3 - 1 - - - - 1 1 - - - - - - - - - - - 11 - 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AH (n=13)</td>
<td>4 4 2 1 2 1 - - - - - - - - - - - - - 1 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF4 (n=12)</td>
<td>1 3 2 - - - 2 - - 1 - 12 - 12 - 12 2 1 - - 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK (n=20)</td>
<td>2 4 2 - - - 20 - - 20 - - - - - - - 9 - 1 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL (n=1)</td>
<td>- - - - - - - - - - - - - - - - - - - - - - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

transition was observed at fixation in AP (n=31).

Four nonsynomous mutations were observed in separate codons in multiple species (codons 947, 967, 999, 1017) (Table 4.4). At codon 947 a T>C transition was observed in one AH allele and one AP allele (codon position 1) and resulted in a phenylalanine to leucine amino acid substitution; at 967 an A>G transition was observed in one AFs.s.s. allele (codon position 2) and one AP allele and resulted in an asparagine to serine substitution; at 999 a T>C transition was observed in two AF4 alleles and one AFs.s.s. allele (codon position 2) and resulted in a valine to alanine substitution; at 1017 a T>C transition was observed in once in AH, AK and AP (codon position 2) resulted in a leucine to proline substitution.

To expand the analysis of sequence variation among the AP complex sibling species, additional VGSC sequence comparisons with available similar coding region sequence fragments were performed with AL (358 bp), AD (213 bp) and AG (348 bp). Between AL and the 90 AP sibling species four synonymous mutations unique to AL were observed (954 T>C, 979 C>T, 981 A>C, 989 G>C, and 1012 C>A). Two synonymous mutations in AL were shared with at least one AP allele (Table 4.4: 995 C>T AP (n=1); 996 C>T AP (n=31), AF4 (n=14)). One synonymous AL mutation occurred at the AP SNP site 981 with a unique sequence change A>C compared to A>G.

Between a partial AD sequence and the AP sibling species alleles eight synonymous mutations unique to AD were identified (978 G>T, 989 G>C [also observed in AL], 993 T>C, 999 C>G, 1004 T>C, 1012 C>A [also observed in AL], 1013
Six synonymous mutations were observed to be shared among one or more AP complex members (956 C>T AH (n=2), AP (n=2); 973 T>C AP (n=1), AK (n=1); 980 T>G AF4 (n=12); 981 A>G AF4 (n=12), 996 C>T AF4 (n=12), AP (n=31); and 1026 C>A AF4 (n=2), AH (n=5), AFs.s. (n=5), AK (n=8), AP (n=10)).

Finally, between a partial AG sequence and the AP sibling species alleles nineteen unique synonymous mutations were observed in AG (940 G>A, 955 T>G, 963 A>G, 969 G>A, 974 T>C, 979 T>C [also observed in AL], 982 G>A, 989 G>A, 992 C>T, 999 C>A, 1002 T>A, 1003 T>A, 1004 C>T, 1008 C>T, 1010 A>G, 1012 C>A [also observed in AL and AD], 1013 C>T [also observed in AD], 1015 G>C, and 1018 T>C).

Nine synonymous SNPs were observed to be shared among one or more AP complex members (956 C>T (also observed in AD) AH (n=2), AP (n=2); 962 C>T AF4 (n=2); 981 A>G (also observed in AD) AF4 (n=12); 983 T>C AP (n=1); 995 C>T (also observed in AL) AP (n=1); 997 C>T AF4 (n=1); 1000 T>C AK (n=1); 1007 A>C AFs.s. (n=11), AP (n=30); and 1026 C>A (also observed in AD) AF4 (n=2), AH (n=5), AFs.s. (n=5), AK (n=8), AP (n=10)).

As described previously, twenty SNPs (Table 4.4; 16 synonymous and 4 nonsynonymous) occurring in the coding region, were found in more than one sequence. Because potential exists for the VGSC gene to be under selection through exposure to insecticides, it becomes necessary to look at the nature of the mutations in the coding regions of the VGSC. The assumption of neutrality was tested by the McDonald-Kreitman test using this coding region data. McDonald-Kreitman null hypothesis assumes the ratio of non-synonymous fixed/synonymous fixed
mutations \((n_F/s_F)\) is equal to non-synonymous polymorphic/synonymous polymorphic mutations \((n_p/s_p)\) (Nei and Kumar 2000). Using a 2 X 2 contingency table, Fisher’s exact test indicated the differences in these ratios among the \(VGSC\) data for AP complex members was non-significant \((p=0.2487)\) indicating no evidence for selection in this region of the gene among the species sampled. However, given that the focus of this analysis was only on a portion of the \(VGSC\) gene, it is possible that selection could be occurring elsewhere, or more observations are needed to visualize selection.

c. Observations of the Introns

As described in Figure 4.1, the \(VGSC\) portion sequenced contained 2 introns. In this dissertation and other published manuscripts these introns are referred to as ‘Intron 1’ and ‘Intron 2’; Intron 1 upstream of the 1014 kdr mutation site and Intron 2 immediately downstream of the 1014 mutation (Martinez-Torres D 1998, Weill et al. 2000, Etang et al. 2009). Intron 1 was of variable length within (exception \(An. farauti\) s.s.) and between species; AP range 880-923 bp, AFs.s.= 928 bp, AH range 939-950 bp, AF4 range 925-938 bp and AK range 940-953 bp. Intron 2 varied in size between species; AP 65 bp, AFs.s. 65 bp, AH 68 bp, AF4 66 bp, and AK 68 bp. Table 4.5 summarizes these differences.

For Intron 1, AFs.s. sequences had the greatest pairwise percent identity, 99.5%. This may not be surprising as all 14 sequences were the same length (928 bp), therefore all differences between AFs.s. members were found to be SNPs. The remaining 4 species were found to have length differences among Intron 1
sequences. Pairwise percent identity of Intron 1 (ranging 880-953 bp) for the remaining within species comparisons were >97.8%: AP (98.1%), AH (97.8%), AF4 (97.8%) and AK (98.1%). Pairwise percent identity was 88.5% when comparing all AP complex members. When comparing all AP sibling species with AL the pairwise percent identity dropped slightly to 87.8%. However, when comparing AL individually to each representative AP complex member, pairwise percent identities were <69% as AL is only represented once in the previous comparison including 90 AP complex representatives, diluting the differences observed between AL and the AP complex.

Intra- (or among species) and inter-species complex relationships are further defined by considering the percent of identical sites within each comparison (Table

<table>
<thead>
<tr>
<th>Intron 1</th>
<th>Aligned Length</th>
<th>Range</th>
<th>Identical Sites (%)</th>
<th>Pairwise % Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>926</td>
<td>880-923</td>
<td>780 (84.2%)</td>
<td>98.1</td>
</tr>
<tr>
<td>AFs.s.</td>
<td>928</td>
<td>928</td>
<td>902 (97.2%)</td>
<td>99.5</td>
</tr>
<tr>
<td>AH</td>
<td>951</td>
<td>939-950</td>
<td>868 (91.3%)</td>
<td>97.8</td>
</tr>
<tr>
<td>AF4</td>
<td>941</td>
<td>925-938</td>
<td>899 (95.5%)</td>
<td>98.8</td>
</tr>
<tr>
<td>AK</td>
<td>955</td>
<td>940-953</td>
<td>872 (91.3%)</td>
<td>98.1</td>
</tr>
<tr>
<td>AP Complex</td>
<td>1007</td>
<td>880-953</td>
<td>521 (51.7%)</td>
<td>88.5</td>
</tr>
<tr>
<td>AP Complex + AL†</td>
<td>1000</td>
<td>880-953</td>
<td>409 (40.9%)</td>
<td>87.8</td>
</tr>
</tbody>
</table>

| Intron 2          | |                  |                    |                     |
|-------------------|----------------|---------------------|---------------------|
| AP                | 65             | 65                  | 63 (96.9%)          | 99.8                |
| AFs.s.            | 65             | 65                  | 63 (96.9%)          | 99.6                |
| AH                | 68             | 68                  | 61 (89.7%)          | 97.7                |
| AF4               | 66             | 66                  | 64 (97%)           | 99.5                |
| AK                | 68             | 68                  | 66 (97.1%)         | 99.1                |
| AP Complex        | 71             | 65-68               | 43 (60.6%)         | 90.8                |
| AP Complex + AL†  | 71             | 64-68               | 31 (43.7%)         | 90.2                |

*AP= An. punctulatus, AFs.s.= An. farauti s.s., AH=An. kinesorum, AF4= An. farauti 4, AK= An. kollensis. § Base pairs. † AL= An. longirostris.
Identical Sites %). Intra-species comparisons observe nucleotide identity over the analyzed gene fragment of 84.2% for AP, 97.2% for AFs.s., 91.3% for AH, 95.5% for AF4, and 91.3% for AK. With inter-species comparisons among AP sibling species allelic percent identity drops to 51.7%; 40% when AL is added to this sequence comparison. This further demonstrates that intronic sequences are most conserved within each species and somewhat conserved between members of the AP complex. However, given that this gene is possibly under selective pressure and natural selection reduces the genetic variation of organisms we would expect to see little to no variation within this gene among species.

Unlike Intron 1, Intron 2 does not vary in size within species and is much smaller (65-68 bp). Pairwise percent identity of all the species was found to be greater than 97.7% (AH) with four species (AP, AFs.s., AF4, and AK) above 99% (Table 4.5). Comparison of all AP complex members showed 90.8% pairwise percent identity and when AL was added to the comparison this percentage dropped slightly to 90.2%. Conserved intronic sequences further demonstrate, as they do in Intron 1, species similarities when analyzing the percent of identical sites. When making intra-species comparisons AP, AFs.s., AF4 and AK each share greater than 96.9% identical sites among themselves, and all AH representatives share 89.7% identical sites. Whereas when comparing all AP complex members together, there are only 60.6% identical sites and 43.7% identical sites when adding AL to the comparison.
Unique species-specific polymorphisms (insertions, deletions and SNPs) contribute to these high similarity percentages within species and lower similarity percentages among species observed within Introns 1 and 2. These differences are demonstrated in Figure 4.2 where the full-length consensus sequence from each of the five An. punctulatus sibling species were aligned by ClustalW as described in Chapter 2. Many unique insertions and deletions specific to each species are observed in this alignment. For example, a 7 bp insertion, in Intron 1 from position 158-165 in the alignment, only appears in An. farauti 4 members and another insertion (7 bp) only occurs in An. hinesorum from position 1100-1106.

2. **Evaluation of Intrinsic Patterns and Simultaneous Monitoring of kdr Mutations**

The observed polymorphic nature of VGSC intronic regions for the sequences described in the previous sections and the tendency of these polymorphisms to appear in a species-specific manner suggested that this gene would be a reliable candidate for molecular species identification. This species identification method could be coupled with probes investigating regions associated with insecticide resistance described previously. As a strategy for species identification, complementary to the ITS2 LDR-FMA described in Chapter 3, targeting the VGSC, this approach would simultaneously monitor resistance-associated mutations in populations currently under insecticide pressure from countrywide distribution of LLINs.
a. **VGSC LDR-FMA multiplex design**

Building from methods to develop LDR-FMA diagnostic assays described in Chapters 2 and 3, species ‘classification’ and the ‘reporter’ LDR probes were designed to hybridize to unique polymorphisms in Intron 1 (AP, AK, AFs.s. and AF4) and Intron 2 (AH) of the PCR amplified VGSC region. Conserved regions of the amplified VGSC surrounding both M918T and L1014F/S mutations allowed for the design of probes to monitor for the presence (heterozygous or homozygous) or absence of these mutations. Locations where kdr mutation detection and species-specific classification LDR reporter probes hybridized with VGSC template PCR products are shaded in the VGSC consensus alignment (Figure 4.2).

The sensitivity and specificity of the species-identification probes was evaluated using PCR products amplified from previously sequenced species-specific controls. Site directed mutagenesis was used to induce the L1014F (TTT) mutation in a cloned representative of the AP complex to serve as a control. This method of mutagenesis however was not successful in creating the L1014S mutation therefore 100 bp oligos were synthesized (IDT Coralville, IA) with sequence complementary to the detection probes (and representative sequenced regions) for wild type (TTA), and both mutants (TTT and TCA) to serve as controls in Figure 4.3. MFI signals from the TTT mutagenesis control and TTT oligo, as well as sequenced AP members known to harbor the wild type TTA and the TTA oligo, fall in the same range (20-25,000).
<table>
<thead>
<tr>
<th>Consensus</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>APConsen</td>
<td>TGTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AKConsen</td>
<td>TGTTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AFssConsen</td>
<td>TGTTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AHConsen</td>
<td>TGTTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AF4Consen</td>
<td>TGTTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>APConsen</td>
<td>TGTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AKConsen</td>
<td>TGTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AFssConsen</td>
<td>TGTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AHConsen</td>
<td>TGTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
<tr>
<td>AF4Consen</td>
<td>TGTTAGCTTT GCGTATATTA TGCACACAG ACTAGACATT TAAGATCTTG TAAAATCTTC AACAATGGCG TTCATGCAGA TTGCACCTCT CTCTCTTACGT TCTC</td>
</tr>
</tbody>
</table>

**Figure 4.2.** Amplified region of Voltage Gated Sodium Channel (VGSC) sequence. Consensus alignment for *Anopheles punctulatus*, *An. koliensis*, An. farauti s.s., An. hinesorum, and An. farauti 4. Consensus sequence for each member was created using *VGSC* sequence from multiple representatives of each species: *An. punctulatus* n=31, *An. koliensis* n=20, An. farauti s.s. n=14, An. hinesorum n=13, An. farauti 4 n=12. Coding region (underlined & italicized) and position of the super-kdr (M918T) and kdr (L1014F, L1014S) mutation sites (*) are defined. Intron 1 and Intron 2 are denoted by the lack of italics and underlines and separate the first and second exons, and second and third exons respectively. LDR probe locations for species determination and kdr detection are denoted by grey (classification probe) and black (reporter probe) coloration. Dashes (-) represent where sequences are in agreement with 'APConsen', locations where nucleotides differ from 'APConsen' are indicated by the insertion of the appropriate nucleotides and gaps in the alignment are represented as full stops (.).
As members of the AP group are known to carry Plasmodium (Burkot et al. 1988, Burkot et al. 1990, Cooper et al. 2009b) and Wuchereria parasites (Bockarie et al. 2002, Benet et al. 2004a) as well as human blood meal (Charlwood et al. 1985, Burkot et al. 1988), 6 additional controls containing Plasmodium falciparum, P. vivax, P. malariae, P. ovale, Wuchereria bancrofti and human genomic DNA were included. For simplicity of illustration the results from the 4 Plasmodium reactions were averaged and included as one data point in Figure 4.3.

![Graph showing specific genomic DNA samples](image)

**Figure 4.3** Detection of species-specific DNAs and kdr mutations by the VGSC Multiplex PCR-LDR-FMA. Data represents a summary of twelve individual control experiments detecting the wild-type (TTA), mutant L1014F (TTT) and mutant L1014S (TCA) kdr mutation sites, as well as differentiating AP, AFs.s., AH, AF4 and AK species. All four human Plasmodium species were tested (averaged data is shown for simplicity) (Plas spp.) as well as Wuchereria bancrofti (Wb) and human (HBS) genomic DNAs specificity examination. Whereas genomic DNAs were added individually into the control reaction, LDR and FMA reactions included oligonucleotide probes representing all five Anopheles species as well as three kdr site probes. Numbers in parentheses next to species designations in the legend identify Luminex FlexMap™ microspheres (Luminex Corp., Austin, TX). 'No DNAs' identifies a blank sample to which no genomic DNAs were added.
Results of the specificity test (Figure 4.3) show by strong allele-specific fluorescent signals (kdr wild type TTA mean fluorescence intensity [MFI] >23,000; kdr mutant TTT >25,000 MFI; kdr mutant TCA >15,000) that each LDR-FMA probe set detected only the kdr allele present in the genomic controls. Given that the TTA and TCA classification probes both terminate in an ‘A’ it is not surprising that we see cross-reactivity MFI around 3000 for these two probes. Simultaneously, each species-specific LDR-FMA probe set detected only the species expected (AP median fluorescence intensity [MFI] >22,000; AK >23,000; AFs.s. >23,000; AH >11,000; AF4 >22,000) and that all species-specific probe background signals were below MFI’s of 2000. This demonstrated that the LDR probes designed to target the species-specific polymorphisms described here detected only the species present and no others. Fluorescent signals from the parasite and human samples were below a background of 160 MFI, demonstrating that the presence of parasites and human blood meal would not obscure Anopheles species identification or kdr mutation detection.

3. Application of VGSC multiplex assay

The VGSC multiplex assay was used to evaluate a collection of 312 mosquitoes for the kdr mutation as well as assign species identification. All 312 mosquitoes were previously assigned species identification using the ITS2 LDR-FMA method described Chapter 2. VGSC species identification was performed without consulting previous ITS2 LDR-FMA species determinations (blind comparisons). VGSC LDR-FMA results demonstrated that all 312 mosquitoes assayed were found to be homozygous for the wild type (sensitive) allele (TTA/TTA) at the kdr mutation
locus. Results in Table 4.6 summarize the comparison of VGSC multiplex assay species identification with the ITS2 LDR-FMA. One discordant sample was detected among the 312 mosquitoes assayed. By VGSC LDR-FMA this mosquito was identified as *An. punctulatus*, where ITS2 LDR-FMA species identified as *An. farauti* s.s.. These results demonstrate 99.7% concordance between the two identification methods as well as the ability of the assay to simultaneously identify the allele present at the kdr mutation locus in a large sample set of field collected mosquitoes.

### Table 4.6

Concordance assessment of molecular ITS2 and VGSC LDR-FMA classification† methods for *An. punctulatus* species complex members

<table>
<thead>
<tr>
<th>ITS2</th>
<th>AP</th>
<th>AFs.s.</th>
<th>AH</th>
<th>AF4</th>
<th>AK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>134</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AFs.s</td>
<td>1</td>
<td>38</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AF2</td>
<td>-</td>
<td>-</td>
<td>47</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AF4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td>AK</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>68</td>
</tr>
</tbody>
</table>

†AP=*An. punctulatus* s.s., AFs.s.=*An. farauti* s.s., AH=*An. hinesorum*, AF4=*An. farauti* 4, and AK=*An. koliensis*

### a. VGSC LDR-FMA analysis of sample set previously subjected to insecticide susceptibility phenotype evaluation

Mosquitoes evaluated for insecticide susceptibility using WHO insecticide bioassays, as described previously, were also analyzed using this assay. Sixty-two mosquitoes, collected from 5 geographically distinct areas (Table 4.1) were previously reported to be susceptible to either 0.05% lambda-cyhalothrin treated filter-paper (n=26, locations: Dimer, Peneng and Lorengau) or new 55 mg/m² deltamethrin-treated LLIN (n=36, locations: Ramu, Naru, and Dimer) (Keven 2010). These mosquitoes were evaluated using the VGSC LDR-FMA. All 62 samples were determined to be homozygous wild type (susceptible) at the kdr locus (TTA;
average MFI 8182) and had minimal background for the mutant resistance probes (TTT and TCA) (average MFI 567). This demonstrates the VGSC LDR-FMA 62 of these samples were both phenotypically susceptible and lacked the mutations associated with knockdown resistance (L1014F/S). Simultaneously, the VGSC LDR-FMA assigned species identification to each of the 62 individual mosquitoes with 100% concordance to previous ITS2 LDR-FMA species identification (Table 4.6) as well as ITS2 RFLP completed by the Entomology Unit of the PNGIMR (Keven 2010). This exercise demonstrated that the VGSC LDR-FMA provides accurate and efficient identification of both AP species composition and kdr mutation status for future studies monitoring insecticide resistance in PNG.

D. Discussion

Here a portion of the Voltage Gated Sodium Channel gene (VGSC) sequence was described for the first time for the AP complex in Papua New Guinea by analyzing 90 individual mosquitoes collected from 17 villages across the country. This gene was chosen for analysis because in other regions of the world, where insecticides are used for pest management (either for agriculture or public health purposes), Anopheles disease vectors, along with many other arthropods, have developed resistance to DDT and pyrethroids, the two most commonly used insecticides. This resistance, termed knockdown resistance, has been most commonly associated with a single point mutation (kdr) in the VGSC, causing a mutation from the wild type leucine at codon 1014 to phenylalanine (L1014F) (Martinez-Torres D 1998, Hemingway and Ranson 2000, Soderlund 2008, Singh et
A variation of the kdr mutation (L1014S) has been described in *An. gambiae* populations and *An. culicifacies* (an Indian malaria vector) (Martinez-Torres D 1998, Hemingway and Ranson 2000, Soderlund 2008, Syafruddin et al. 2010). Additional mutations associated with insecticide resistance have been shown to occur in the region surrounding kdr for other arthropods, but have not been detected in *Anopheles* species to date (Soderlund 2008).

PNG has had a long and variable history of insecticide usage. Because LLINs impregnated with pyrethroids (PermaNet®, Vestergaard Frandsen, Lausanne, Switzerland; treated with permethrin) are currently being distributed to every household in the country, monitoring for development of resistance to these compounds is important for vector management success. Of note, a study from neighboring Indonesia, which has experienced vector control efforts similar to PNG over the past 60 years (Baird et al. 1996), has recently reported the identification of the kdr mutant allele. In this study the kdr mutant allele (TTT) has recently been detected among four major malaria vectors, *An. sundiacus*, *An. subpictus*, *An. vagus* and *An. aconitus* in 55.8% of the surveyed population (Syafruddin et al. 2010).

Features of the *VGSC* gene that make it a unique target for species evaluation include juxtaposition of intron and exon sequence, and containment of kdr SNPs associated with insecticide resistance. Therefore sequencing this region serves multiple purposes for this dissertation: monitoring for the development of the kdr mutation(s), testing conservation of the coding sequence and capturing sequence variation within the introns, useful to address questions of species relationships.
Therefore analysis of this gene provides additional insight, beyond analysis of ITS2 regarding the integrity of AP species complex diversity and sibling species relationships. Observation that intron sequence polymorphisms could be used to produce consistent species differentiations (similar to that observed for ITS2) provides evidence that species specific sequence divergence has occurred at multiple regions in AP sibling species genomes.

Regions of interest for kdr mutations were captured in the length of the sequence and all 90 individuals were found to harbor the wild type sensitive alleles at the super-kdr (M918T) and kdr (L1014F/S) loci. Table 3.2 and Table 4.2 demonstrate that when representatives of the same species are compared among themselves >98.4% pairwise percent identity is observed for both ITS2 and VGSC. For VGSC Intron 1, length differences were observed in AP, AH, AF4 and AK. All 14 members of AFs.s. did not have length polymorphisms in Intron 1. An. farauti s.s. was also observed to have the greatest PPI (99.4%) even though these representative mosquitoes were collected from geographically distinct villages in 2 mainland provinces (MADANG: Bilbil, Dimer, Hudini, Sausi MOROBE: Godowa,) one collection from Manus Province (Lorengau), an island >340 km from mainland PNG.

Comparisons of the coding region of the VGSC demonstrated homology (>80.3% identical shared sites, >97.8% PPI) between members of the AP complex and reference sequences available for An. dirus and An. gambiae as well as an An. longirostris representative sequenced in this project (Table 4.3). Is it observable, however, that An. longirostris shares a similar amount of identical sites with the
entire AP complex as all AP complex members share among themselves (301/358 [84.1%] and 305/358 [85.2%] respectively). This is not the case when observing comparisons between the AP complex and An. dirus or An. gambiae as they both only share 80.3% of identical sites with the AP complex. This comparison exercise was thought to possibly elucidate relationships between AP complex members and related mosquitoes that are thought to be close (An. longirostris), moderate (An. dirus) and distant (An. gambiae) relatives. The results here, however, do not strongly demonstrate these relationships since similar results are observed between the AP complex comparisons with An. dirus and An. gambiae.

A subset of samples (n=62) were previously evaluated for phenotypic insecticide susceptibility as well as species identification by ITS2 RFLP (Keven 2010) and ITS2 LDR-FMA. VGSC LDR-FMA evaluation all 62 samples were found to be homozygous wild type at the kdr locus and species identification were in 100% agreement with the two ITS2 species identification methods previously performed on the sample set. This exercise demonstrated the application of the assay to a field survey of insecticide susceptibility.

Since it is likely that the VGSC or portions of the VGSC could have been subjected to selection from past insecticide usage, the coding region synonymous and non-synonymous mutations were used to test the assumption of neutrality using the McDonald-Kreitman test. This test indicated the differences between fixed and non-fixed synonymous and non-synonymous were not significant (p=0.2487)
therefore there was no evidence of selection occurring in this region of the gene for these representatives.

Intronic sequence evaluation revealed polymorphic regions of Intron 1 and Intron 2. Unique species-specific insertions and deletions were observed across the entire length of both introns. The availability and consistency of these insertion/deletion polymorphisms proved to be useful for simultaneous diagnosis of AP species and kdr-associated mutations. Additionally, these features of intron sequence variation provide further evidence that species-specific divergence has occurred at multiple locations in the AP sibling species genomes.

Prior to the *VGSC LDR-FMA* method, DNA sequencing was the only way to identify the nucleotide at the kdr locus and the zygosity of this mutation. It is possible to differentiate ‘TTT’ from ‘TTA’ at L1014F for AP complex species using a restriction enzyme (Keven 2010), however this method is limited and will not distinguish the ‘TCA’ mutation from the wild type ‘TTA’ and determining the zygosity of the locus is difficult. So prior to the *VGSC LDR-FMA* method, to assign a species identification as well as identify if the ‘TTT’ mutation was present, two separate PCR-RFLPs would have to be conducted on each sample. The *VGSC LDR-FMA* method reduces the task to one PCR and LDR-FMA and alleviates waste from ethidium bromide and the need to interpret two agarose gel RFLPs, while accommodating a more high-throughput approach with more informative and through results in a platform that is conveniently expandable.
The evaluation of the sample set (n=312) described in this chapter suggests that the polymorphisms targeted for species identification are robust and occur in populations that may not have the frequent opportunity to exchange genetic information. These same results were observed with the polymorphisms targeted in Chapter 3 with the ITS2 sequence, adding support for the existence of independent sibling species. Because of multiple lines of evidence for independent segregation of sequence polymorphisms within and not between the AP sibling species, it can be hypothesized that insecticide resistance will evolve independently for each species. Therefore each individual species must be monitored for evolution of insecticide resistance.

Point mutations observed to evolve at other loci, such as acetylcholinesterase (AChE) genes and GABA receptor genes, have been found in mosquito and insect species following insecticide pressure. Metabolic insecticide resistance in insects through over-expression of detoxifying enzymes (cytochrome p450s, carboxylesterases, glutathione S-transferases) have also been reported in mosquitoes (Hemingway et al. 2004) and can co-occur with resistance associated SNPs like kdr. Since resistance has been shown to develop through different avenues and in combination, monitoring more than one locus following insecticide exposure is optimal. Therefore future work with this project will yield additional monitoring tools to meet this need.
V. CHAPTER 5

Evaluating Diversity of the *Anopheles punctulatus* species Complex by Investigating Mitochondrial Genes
A. Introduction: Mitochondrial DNA- Advantages, disadvantages, and what is known about the AP complex mitochondrial sequence

The circular molecule of mitochondrial DNA (mtDNA) is maternally inherited in most eukaryotic organisms including insects. Because of this inheritance pattern, and because the mitochondrial genome lacks recombination and has a high rate of evolution (~5-10 times that of single copy nuclear genes), it is particularly useful in genetic comparison studies of recently evolved/evolving or closely related species (Avise 1987, Hartl and Clark 1997). To date, complete mitochondrial genomes exist for four *Anopheles* species belonging to subgenera *Cellia* (African malaria vectors *An. gambiae* (Beard et al. 1993) and *An. funestus* (Krzywinski et al. 2006)), *Anopheles* (*An. quadrimaculatus*, a North American species (Mitchell et al. 1993)), and *Nyssorhynchus* (*An. darlingi*, a Central and South American species (Moreno et al. 2010)) (Figure 1.2). These mitochondrial genomes have been used to estimate lineage divergence times, evaluate taxon differentiation and detect incipient speciation. Additionally these reference genomes serve as a helpful guide to study the mtDNA of other *Anopheles* species.

Because mtDNA lacks recombination, in theory any mitochondrial gene will demonstrate a single evolutionary path. However, individual genes within the mitochondria are thought to evolve at different rates (Moritz et al. 1987). Because of these characteristics, arguments have been made, both supporting and refuting, that mtDNA provides the most accurate depiction of species relationships (Moore 1995, Hoelzer 1997, Moore 1997). Using mtDNA for genetic comparisons does have some caveats, as there is no ‘gold standard’ mitochondrial or nuclear gene for assessing
evolutionary relationships. For example due to mtDNA inheritance pattern, in small populations the loss/gain of mtDNA haplotypes will be great. The fixation of an mtDNA lineage only requires small amounts of migration between populations. Additionally, technical issues can arise when amplifying mtDNA genes (Roderick 1996). Within an organism there can be copies of pseudogenes, pieces of mitochondrial sequences that have integrated into nuclear DNA, as well as symbiotic organisms with similar genetic sequences (Roderick 1996, Zhang and Hewitt 1996, Kittayapong et al. 2000).

Nevertheless, whole mitochondrial genome analysis of Anopheles species has allowed researchers to approximate the split of Anopheles and Aedes/Culex (Anophelinae and Culicinae subfamilies) (145-200 Mya) (Figure 1.2) as well as relationships within the Anopheles genus where the split between subgenera Anopheles and Cellia has been dated between 58-106 Mya (Krzywinski et al. 2006, Moreno et al. 2010). Further speciation within Cellia for two African malaria vectors, An. gambiae and An. funestus has been estimated between 36-80 Mya (Krzywinski et al. 2006, Moreno et al. 2010). Sequencing of these genomes and comparison with other Dipteran mitochondrial genomes has revealed Anopheles mitochondrial genomes to be 80% similar to Drosophila and approximately 90% similar to one another, with conserved gene order among the four Anopheles. Anopheles gene order differs slightly from Drosophila where a strand of tRNAs (arginine [R], alanine [A], asparagine [N], serine [S], glutamate [E] and phenylalanine [F]) are arranged in a different order in Drosophila (F, E, S, N, R, A) but maintain the same location in genome. Specifically, the highest degree of
similarity between genomes was found within the Cytochrome Oxidase I (COI) sequence with >90% similarity observed between Anopheles species and Drosophila at the protein sequence level (>80% at the DNA sequence level).

Mitochondrial DNA sequence analysis for members of the AP complex has been reported in a study focused on Cytochrome Oxidase II (COII). In an attempt to reconstruct the phylogeny of mosquitoes in the Australasian region, a 684 bp region of COII was compared among individuals from the AP complex (An. punctulatus, An. spp. near punctulatus, An. koliensis, An. farauti s.s., An. hinesorum, An. torresiensis, An. farauti 4-6, and An. irenicus) as well as non-AP complex mosquitoes found in the region. This comparison, which utilized a single COII sequence to represent each species, suggested that the AP complex is monophyletic. However, relationships among members of the AP complex were unresolved due to low BST support for branches within the phylogenetic tree and the authors suggested that further sequencing of larger portions of the mtDNA coupled with nuclear sequences might help improve resolution (Foley et al. 1998). No further comparisons have been published for the AP complex using COII or any other mitochondrial genes.

At the time this project began, the complete mtDNA of the AP complex was not known. A goal of the project was to assess inter- and intra- species diversity and relationships; therefore this chapter of my dissertation will focus on individual mitochondrial genes for genetic comparison purposes.

Since multiple copies of the mtDNA exist within a eukaryotic cell and mtDNA lack large introns potentially inhibitory to amplification and sequencing, and because the previously mentioned mitochondrial genomes exist for other Anopheles,
PCR amplification (Simon et al. 1994) and DNA sequencing is thought to be relatively straightforward. Therefore two mitochondrial genes were chosen to assess the multiple species hypothesis by quantifying diversity within and among the five AP complex representatives that are the focus of this dissertation.

_COI_ was chosen for mtDNA analysis of the AP complex. This gene is of great interest for many reasons. The Consortium of the Barcode of Life (CBOL) has targeted _COI_ as a gene of choice and seeks to establish a database with the _COI_ gene from as many organisms as possible to further organism-, population- and community-level research (Ratnasingham and Hebert 2007). Additionally, this gene has been used extensively in evolutionary analyses of distant and closely related Dipteran species, including other cryptic _Anopheles_ species (Lunt et al. 1996, de Brito et al. 2002, Dusfour et al. 2004, Foley et al. 2007), but has never been used to address questions within the AP complex. As previously mentioned, mtDNA genes have been shown to evolve at different rates. The mutation rate of _COI_ is not thought to fall on either extreme, as the gene has a mixture of conserved and variable regions (Lunt et al. 1996). Thus, _COI_ provides an alternative genetic perspective to the two nuclear sequences discussed previously in Chapter 3 and 4 and will contribute to a better understanding of the mtDNA in the AP complex.

Since _COI_ is thought to be useful in differentiating closely related species, studies with other cryptic species groups have utilized _COI_ to demonstrate instances of gene flow, detect introgression and establish between-species relationship patterns (Besansky et al. 1997, Walton et al. 2000). For example, a population evaluation of _An. dirus_ cryptic species group amassed 70 distinct haplotypes for _COI_.

127
An. dirus, is taxonomically located, along with the AP complex, in the Neomyzomyia series of the Cellia subgenus. As a South East Asian mosquito, it would be hypothesized that An. dirus is a relative of the AP complex (Beebe et al. 1999b). A cryptic species group, An. dirus is thought to be composed of at least 7 morphologically indistinguishable species (Walton et al. 1999, Morgan et al. 2010), much like the AP sibling species group.

Walton and colleagues focused on three members of the Dirus complex: An. dirus A, C and D. Distribution patterns of these three species have been shown to be distinct but overlapping. An. dirus D is found predominantly in Bangladesh, Eastern India, and Myanmar. An. dirus A is found primarily in Vietnam, Laos, Cambodia and Thailand. An. dirus A and D are found in sympatry along the Myanmar-Thailand border. An. dirus C has more specific habitat preferences and is found almost exclusively near limestone karst landscapes located in small areas in southeast Thailand and along the Thai peninsula. In these areas An. dirus C has been observed to reside sympatrically with either An. dirus A or D.

In their COI-based assessment of population structure, 84 individual mosquitoes were collected from 11 sites within Bangladesh, Myanmar and Thailand (Walton et al. 2000, Walton et al. 2001). From these 84 individuals, 70 COI haplotypes were observed (a similar region of COI as used in this dissertation). Four haplotypes were shared between An. dirus A (n=4) and D (n=4). With the exception of one haplotype, the originating mosquitoes were collected from geographically different collection sites either within Myanmar or Thailand. An. dirus C did not share any COI haplotypes with An. dirus A or D (Walton et al. 2000). Therefore,
results of Walton et al. demonstrate the usefulness of mtDNA in speciation and phylogeography research.

To augment the data generated for COI among the AP complex, an additional mitochondrial gene was chosen. Cytochrome B (Cytb) has been shown to supplement COI analysis, as well as the nuclear sequence ITS2, in studies investigating cryptic Anopheles species groups (Besansky et al. 1997, Mukabayire et al. 1999, Krzywinski et al. 2001a, Dusfour et al. 2004). Because there is no recombination in mtDNA, relationships established using Cytb are expected to be identical to COI. Using these two mitochondrial genes will not only complement the nuclear DNA analysis described in Chapters 3 and 4, but also allow for intra- AP complex species comparisons and provide a framework from which future whole mitochondrial genome sequencing can be assembled and compared. I hypothesize that these mitochondrial genes will harbor species-specific polymorphisms similar to observations within the nuclear genes previously analyzed (ITS2 and VGSC) supporting the multiple species hypothesis.

B. Methods

Details regarding sampling methods, DNA extraction, amplification of COI and Cytb and sequencing can be found in Chapter 2. DNA sequencing of COI and Cytb was completed for 97 and 37 individual mosquitoes, respectively (Table 5.1, Figure 2.2). Sequence assembly and analysis methods are described in Chapter 2.
C. Results

1. Sequencing of mtDNA for *An. punctulatus* Sibling Species

DNA sequence analysis was performed for 103 individual mosquitoes for *COI* (n=66), *Cytb* (n=6) or both (n=31) (Table 5.1). Sequencing efforts were focused on the five species previously discussed in this dissertation (AP, AFs.s., AH, AF4 and AK). Individual mosquitoes were subjected to PCR amplification and DNA sequencing of a 1270 bp portion of *COI* (n=97) and a 470 bp portion of *Cytb* (n=37) using adapted methods (Simon et al. 1994, Lunt et al. 1996, Krzywinski et al. 2001a, Dusfour et al. 2004) described in Chapter 2 (Table 2.2).

<table>
<thead>
<tr>
<th>Site</th>
<th>Province</th>
<th>Species Evaluated*</th>
<th>P</th>
<th>F1</th>
<th>H</th>
<th>F4</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albulum</td>
<td>EAST SEPIK</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dreikikir</td>
<td>EAST SEPIK</td>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nale</td>
<td>EAST SEPIK</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Nanaha</td>
<td>EAST SEPIK</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peneng</td>
<td>EAST SEPIK</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yawatong†</td>
<td>EAST SEPIK</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilbil</td>
<td>MADANG</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimer</td>
<td>MADANG</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hudini</td>
<td>MADANG</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Maraga</td>
<td>MADANG</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuru†</td>
<td>MADANG</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sausi†‡</td>
<td>MADANG</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yagaum</td>
<td>MADANG</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Lorengau</td>
<td>MANUS</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gingala†</td>
<td>MOROBE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Godowa†</td>
<td>MOROBE</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Mumeng</td>
<td>MOROBE</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ramu</td>
<td>MOROBE</td>
<td>9</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitapena</td>
<td>WEST SEPIK</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Yapsie†‡</td>
<td>WEST SEPIK</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kuru</td>
<td>WESTERN</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singopra</td>
<td>WHP**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

*Number of individual mosquitoes included in this study; P= *An. punctulatus*, F1= *An. farauti* s.s., H= *An. hinesorum*, F4= *An. farauti* 4, K= *An. koliensis*. †*An. longirostris* collected in these locations for *COI* genetic comparison purpose. ‡*An. longirostris* collected in these locations for *Cytb* genetic comparison purposes. **WHP= Western Highlands Province
According to the *An. gambiae* mtDNA reference genome (Genbank Accession no. NC00284) the COI gene is 1537 bp in length encoding 512 amino acids (Beard et al. 1993). After trialing several primer combinations the most successful amplification method employed for COI captured a region from nucleotide 267-1536 in the *An. gambiae* mtDNA reference, starting at the third position of codon 89 and continuing through codon 512: AP (n=32, Genbank accession nos. HQ840857-HQ840888), AFs.s. (n=17; HQ840792-HQ840808), AH (n=13; HQ840809-HQ8408821), AF4 (n=16; HQ840822-HQ840837) and AK (n=19; HQ840838-HQ840856). Additionally, an *An. longirostris* (HQ840889) representative was included for PCR and DNA sequencing of this region for genetic comparison purposes.

*Cytb* of *An. gambiae* is 1137 bp in length and encodes 378 amino acids (Beard et al. 1993). After trialing different primer combinations the most successful amplification method employed captured a region (470 bp) that spanned *An. gambiae* reference nucleotide 409-878: codon 137 to the second nucleotide of codon 293. A total of 37 sequences were captured for this region, 31 of which had previously been sequenced for COI: AP (n=6, Genbank accession nos. HQ840921-HQ840926), AFs.s. (n=4; HQ840890-HQ840893), AH (n=18; HQ840894-HQ8408911), AF4 (n=6; HQ840912-HQ840917) and AK (n=3; HQ840918-HQ840920). *An. longirostris* (HQ840927-HQ840928) representatives were also included for PCR and DNA sequencing of *Cytb* for genetic comparison purposes.
Sequences were inspected for similarity to the mtDNA genomes of *An. gambiae*, *An. funestus*, *An. darling* and *An. quadrimaculatus*. These comparisons were made to assess potential that the sequenced products were possible nuclear pseudogenes (Zhang and Hewitt 1996) or mtDNA-like genes of *Wolbachia*, a common endosymbiont of many insects including *Anopheles* (Kittayapong et al. 2000). *COI* Consensus sequences for each species were compared for the five AP siblings and with the reference mtDNA genomes (Table 5.2).

**Table 5.2 Pairwise Percent Identity* of mtDNA sequences among *Anopheles*** species**

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AFs.s.</th>
<th>AH</th>
<th>AF4</th>
<th>AK</th>
<th>AG</th>
<th>AFun</th>
<th>ADar</th>
<th>AQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFs.s.</td>
<td>93.1</td>
<td>94.7</td>
<td>93.2</td>
<td>96.9</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AH</td>
<td>94.4</td>
<td>96.3</td>
<td>91.2</td>
<td>93.2</td>
<td>92.1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF4</td>
<td>94.0</td>
<td>94.0</td>
<td>93.7</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK</td>
<td>92.6</td>
<td>93.6</td>
<td>93.5</td>
<td>90.8</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td>88.3</td>
<td>89.0</td>
<td>88.9</td>
<td>87.0</td>
<td>88.8</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFun</td>
<td>87.7</td>
<td>88.7</td>
<td>88.0</td>
<td>88.0</td>
<td>87.2</td>
<td>87.7</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADar</td>
<td>89.1</td>
<td>90.4</td>
<td>90.0</td>
<td>89.3</td>
<td>89.8</td>
<td>88.0</td>
<td>90.6</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AQ</td>
<td>87.7</td>
<td>87.9</td>
<td>88.7</td>
<td>87.2</td>
<td>88.0</td>
<td>86.6</td>
<td>88.3</td>
<td>88.0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>91.7</td>
<td>90.0</td>
<td>91.0</td>
<td>91.5</td>
<td>91.5</td>
<td>89.1</td>
<td>n.d.</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

*values displayed are percentages. **AP complex species as well as *An. gambiae* (AG), *An. funestus* (AFun), *An. darling* (ADar) and *An. quadrimaculatus* (AQ). Unshaded values correspond to *COI* comparisons, shaded italicized values correspond to *Cytb*. n.d.=not determined as most of the *An. funestus* *Cytb* gene is not complete.

a. **General mtDNA sequence observations**

When creating alignments of both mtDNA sequences it became clear from general visual observation, that within most species, individual sequences existed
appearing to be very different from the rest of the members of that species. Figure 5.1 A demonstrates this observation from the 17 An. farauti s.s. COI sequences where sequence numbers 16 and 17 display significantly more of polymorphisms when compared to the other 15 AFs.s. sequences (black bars indicate regions of differences between the majority). Contrasting these results with the other AP siblings no apparent extreme differences were observed in the AK group of sequences (Figure 5.1 B). These sequences in AFs.s., which appeared to be very different from the rest of the species representatives, were removed from the sequence alignment and labeled as ‘outliers’. Additionally, within An. hinesorum Cytb alignment it was clear that there were distinct groups of sequences, therefore this alignment was divided into two groups, Group A and Group B (Figure 5.2). Two outliers were also recognized and pruned before analyses, removing them from the alignments. Consensus sequences were then created using the total and ‘pruned’ alignments for both sets of mtDNA sequences. Summary statistics comparing all alignments are presented in Tables 5.3 (COI) and 5.4 (Cytb) comparing the alignments both before (total) and after outlier sequences were removed (pruned). It is important to mention that sequences removed from the alignments were not discarded and their characteristics will be discussed further in later portions of this section.

After outlier sequences were removed from the alignments, pairwise percent identity and identical sites increased for both COI and Cytb alignments. For example, by removing 3 outliers from the COI AF4 alignment PPI increased 1.6% (97.9%>99.5%) and identical sites increased>10% from 1099/1270 to 1235/1270
Figure 5.1 Example of observed outliers in mtDNA sequence analysis. Screenshots of whole sequence alignments in Geneious of AFs.s. (A.) and AK (B.) COI sequence (1270 bp) in Geneious demonstrate where sequences differ from the majority (black bars). Two clear outliers exist in the AFs.s. alignment (sequence #16 and #17). These two sequences were removed (during pruning) from consensus creation and designated as outliers. Additionally, less obvious differences between sequence #12-15 (black bracket) and the rest of the alignment are also observable. These four sequences happen to be from Manus Province indicating a phylogeographical difference among these samples. It does not appear any sequences in the AK alignment are substantially different as observed in the AFs.s. alignment, therefore no sequences were designated as outliers and removed from the AK alignment.
(Table 5.3). Additionally by removing 2 outliers from the Cytb AP alignment PPI increased >4% (95.0%>99.8%) and identical sites increased >10% from 419/470bp to 468/470bp (Table 5.4). This demonstrated that those sequences removed were, in fact, different from the remaining group and impacting the overall intra-species sequence comparison statistics of percent similarity and shared identical sites.

There was no evidence of COI or Cytb sequence outliers for An. koliensis. An. farauti s.s. and An. farauti 4 did not appear to have any obvious outliers in Cytb alignments.

In Tables 5.3 and 5.4, once consensus sequences had been created after pruning outliers the revised alignments were compared with An. longirostris and An. gambiae reference sequences. Identical sites shared between the compared groups are displayed as percentages. Observed differences between the consensus from each species and the outliers produce a similar percentage of shared sites as comparisons of the consensus with the two outgroups.
### Table 5.3 COI* Sequencing Results

<table>
<thead>
<tr>
<th>Species</th>
<th>n</th>
<th>Pairwise % Identity (PPI)</th>
<th>Identical Sites**</th>
<th>Identical sites for Consensus¹ vs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Total</td>
<td>Pruned</td>
<td>Total</td>
</tr>
<tr>
<td>An. punctulatus</td>
<td>32</td>
<td>29</td>
<td>98.1%</td>
<td>99.7%</td>
</tr>
<tr>
<td>An. farauti s.s.</td>
<td>17</td>
<td>15</td>
<td>97.9%</td>
<td>99.3%</td>
</tr>
<tr>
<td>An. hinesorum</td>
<td>13</td>
<td>11</td>
<td>98.2%</td>
<td>99.6%</td>
</tr>
<tr>
<td>An. farauti 4</td>
<td>16</td>
<td>13</td>
<td>97.9%</td>
<td>99.5%</td>
</tr>
<tr>
<td>An. koliensis</td>
<td>19</td>
<td>19</td>
<td>98.7%</td>
<td>98.7%</td>
</tr>
</tbody>
</table>

*1270 bp, **nucleotides, †Sequences remaining after pruning, ¹Consensus sequence for each species derived from pruned alignment, ²sequences removed in 'pruning', ³An. longirostris (AL) reference sequence (HQ840889), ⁴An. gambiae mtDNA reference (NC002084)

### Table 5.4 Cytb* Sequencing Results

<table>
<thead>
<tr>
<th>Species</th>
<th>n</th>
<th>Pairwise % Identity (PPI)</th>
<th>Identical Sites**</th>
<th>Identical sites for Consensus¹ vs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Total</td>
<td>Pruned</td>
<td>Total</td>
</tr>
<tr>
<td>An. punctulatus</td>
<td>6</td>
<td>4</td>
<td>95.0%</td>
<td>99.8%</td>
</tr>
<tr>
<td>An. farauti s.s.</td>
<td>4</td>
<td>4</td>
<td>99.1%</td>
<td>99.1%</td>
</tr>
<tr>
<td>An. hinesorum A</td>
<td>18</td>
<td>12</td>
<td>96.3%</td>
<td>99.6%</td>
</tr>
<tr>
<td>An. hinesorum B</td>
<td>18</td>
<td>4</td>
<td>97.6%</td>
<td>97.6%</td>
</tr>
<tr>
<td>An. farauti 4</td>
<td>6</td>
<td>6</td>
<td>99.6%</td>
<td>99.6%</td>
</tr>
<tr>
<td>An. koliensis</td>
<td>3</td>
<td>3</td>
<td>98.3%</td>
<td>98.3%</td>
</tr>
</tbody>
</table>

*470 bp, **nucleotides, †Sequences remaining after pruning, ¹Consensus sequence for each species derived from pruned alignment, ²sequences removed in 'pruning', ³An. longirostris (AL) reference sequence (HQ8408927-28), ⁴An. gambiae mtDNA reference (NC002084)
b. **COI diversity within the AP complex**

Haplotypes were established using the pruned sequence (n=87) alignments in DnaSP v5 (Librado and Rozas 2009) which allowed for the calculation of heterozygosity ($H$; defined as haplotype diversity) and the nucleotide diversity per site ($\pi$). $H$ values range from 0-1, with 0 indicating no diversity (all sequences exactly the same) and 1 indicative of diversity (no shared polymorphisms). These results (Table 5.5) highlighted the existence of 76 total haplotypes among the 87 sequences subjected to analysis. $H$ values were all > 0.89 indicating diversity within each species. For example, AK had an $H$ value of 1.000 as all 19 AK sequences included for analysis harbored a unique haplotype. Even though $H$ values are high for each species, the $\pi$ values for are low, indicating very few nucleotide changes exist per site, which is also reflected in the high PPI and identical sites in Table 5.3. The pruned sequence alignment was used in this calculation to demonstrate total haplotypes among the sequences that appeared to be the same species.

All but 17 sequences differed at least at one nucleotide position in the 1270 bp analyzed. These 17 sequences were included in one of the 6 haplotypes: AP contained three haplotypes (one haplotype is shared by a representative from

<table>
<thead>
<tr>
<th>Species</th>
<th>n</th>
<th>h</th>
<th>S</th>
<th>$H$</th>
<th>$\pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>29</td>
<td>23</td>
<td>48</td>
<td>0.975</td>
<td>0.00390</td>
</tr>
<tr>
<td>AFs.s.</td>
<td>15</td>
<td>14</td>
<td>36</td>
<td>0.990</td>
<td>0.00691</td>
</tr>
<tr>
<td>AH</td>
<td>11</td>
<td>8</td>
<td>21</td>
<td>0.891</td>
<td>0.00843</td>
</tr>
<tr>
<td>AF4</td>
<td>13</td>
<td>12</td>
<td>35</td>
<td>0.987</td>
<td>0.00125</td>
</tr>
<tr>
<td>AK</td>
<td>19</td>
<td>19</td>
<td>91</td>
<td>1.000</td>
<td>0.01321</td>
</tr>
<tr>
<td>Total</td>
<td>87</td>
<td>76</td>
<td>300</td>
<td>0.995</td>
<td>0.05971</td>
</tr>
</tbody>
</table>

\[1\] Cytochrome Oxidase I gene; 1270bp, n=number of sequences, h=haplotypes, S=number of polymorphic sites, $H$=heterozygosity, $\pi$=nucleotide diversity per site
Mumeng [Morobe Province] [Table 5.1] and Pitapena [West Sepik Province], another is shared by representatives [n=3] from Albulum and Yawatong [East Sepik], and Mumeng [Morobe] and lastly, the third haplotype is shared by three representatives from Yagaum [Madang] and one from Dreikikir [East Sepik]. AFs.s. contained one haplotype (shared by two individuals from Dimer [Madang]), AF4 also contained one haplotype (shared by two individuals from Dimer) and finally, AH contained a haplotype (shared by four individuals, all from the Ramu collection site [Morobe]).

These same calculations were generated for the total sequence alignment where all 97 sequences were included. Nine new haplotypes were recognized from the additional 10 sequences that had been pruned and considered to be outliers. All but one outlier sequence was a unique haplotype, a single AFs.s. representative shared a haplotype with an AK sequence. The H of all the species remained similar to that of the pruned alignment (>0.923).

c. Closer examination of COI sequences: sequence outliers

Closer inspection of the COI sequence was performed for the outliers that had been pruned out of the sequence alignments. These 10 sequences were from individual mosquitoes previously identified by more than one method: ITS2 LDR-FMA, VGSC LDR-FMA, ITS2 RFLP and/or ITS2 and VGSC sequence comparisons. In comparing their COI sequence with the consensus sequences created from each species, 8 of the 10 appeared to associate with a species different from their previously determined species identification. Figure 5.3 demonstrates this
observation in a portion of the COI alignment (160 bp out of 1270 bp). In this figure outliers are aligned with the species consensus to which they are most similar. For instance, AF4_hud_05, an individual previously identified as AF4 from the Hudini collection site, and AH_nal_01, an individual previously identified as AH from the Nale collection site, are most similar to the AP consensus sequence (both share 1265/1270 sites [99.6%]), and do not share homology with the AF4 or AH consensus. In the alignment, sequence #13 (AF1_mad_14) an individual identified as AFs.s. from Dimer, is most closely associated with the AK consensus (1260/1270 [99.2%]). This sequence actually shares a haplotype with an AK individual (AK_mad_13) also collected from Dimer, both having the exact same COI sequence. An. hisesorum consensus is the only representative that does not have an associated outlier.

Two sequences do not appear to be
closely associated with a consensus from any of the five AP species: AH_wip_01, an individual previously identified as AH from Kuru, and AP_fin_01, an individual previously identified as AP from Godowa. These sequences share some positions in common with the consensus sequences but, overall, they are largely different (1076/1270 [84.7%] and 1042/1270 [82%] respectively).

These 10 sequences were analyzed further to characterize their potential origin. None of the 10 sequences harbored premature stop codons or gaps to shift the reading frame. To further ensure these COI sequences were Anopheles in origin the Basic Local Alignment Search Tool (BLAST) megablast suite (http://blast.ncbi.nlm.nih.gov/) was used to query Genbank for similar nucleotide sequences. Querying the entire nucleotide collection of Genbank, with no restrictions set, the first 100 hits were from Anopheles species. This demonstrated that these two COI sequence highly similar to other Anopheles and are probably of Anopheles origin. However it is also possible that mosquito storage protocols could have contributed to these observed outlier mtDNA sequences that associated with different species. In the collection and morphological identification process, mosquitoes that are assigned the same morphological identification are occasionally stored in a single tube prior to DNA extraction. However, many of the individuals analyzed in this chapter for COI were also analyzed in previous chapters for other genes and these incongruent patterns were not observed. Since this project is the first to report on the COI sequences of AP complex members, it is unclear whether these outliers may be similar to other AP complex members not included in this dissertation.
Because of the observations in this section an effort was made to ensure that the samples that associated with different *COI* consensus sequences, were truly representative of those individuals and not a product of laboratory or sequencing error. Therefore these samples were subjected to a second round of PCR amplification and DNA sequencing for this gene as well as ITS2, and produced the sequences congruent with the previous results.

**d. Diversity of *Cytb* sequence as a complementary mtDNA sequence**

Using the pruned alignments for *Cytb* discussed in previous sections, haplotypes (h), variable sites (S) and heterozygosity (*H*) were calculated (Table 5.6) in the same manner as *COI*. *H* values were all > 0.83 indicating diversity within and among the species reflecting that almost all individuals of each species were of a unique haplotype (similar to what was observed with *COI*). Haplotypes shared by more than one individual were only observed for AP, AFs.s. and AH group A sequence groups. AFs.s. had one haplotype shared by two individuals both from Lorengau (Manus Province). AP had one haplotype shared by two individuals, from Albulum and Peneng (East Sepik). AH group A has three haplotypes shared by more than one individual.

<table>
<thead>
<tr>
<th>Species</th>
<th>n</th>
<th>h</th>
<th>S</th>
<th><em>H</em></th>
<th><em>π</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0.833</td>
<td>0.00213</td>
</tr>
<tr>
<td>AFs.s.</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>0.833</td>
<td>0.00851</td>
</tr>
<tr>
<td>AH group A</td>
<td>12</td>
<td>7</td>
<td>7</td>
<td>0.894</td>
<td>0.00394</td>
</tr>
<tr>
<td>AH group B</td>
<td>4</td>
<td>4</td>
<td>22</td>
<td>1.000</td>
<td>0.02376</td>
</tr>
<tr>
<td>AF4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>1.000</td>
<td>0.00426</td>
</tr>
<tr>
<td>AK</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>1.000</td>
<td>0.01707</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>26</td>
<td>74</td>
<td>0.983</td>
<td>0.04388</td>
</tr>
</tbody>
</table>

1 *Cytochrome b* gene; 470 bp, n=number of sequences, h=haplotypes, S=number of polymorphic sites, *H*=heterozygosity, *π*=nucleotide diversity per site.
one shared by two individuals the other by three individuals, were composed of sequences from the Ramu (Morobe) collection site only. The third haplotype was shared by two individuals from Ramu and one from Nale (East Sepik).

Outlier sequences (n=4) were compared with consensus sequences generated for each species. Two of these Cytb outliers were also COI outlier sequences: AH_nal_02 and AH_wip_01. AH_nal_02 most closely associated with AP COI sequence and when compared with Cytb reference sequences also associated with the AP consensus. AH_wip_01, previously, did not closely associate with any AP complex member consensus (Figure 5.3). In the case of Cytb this sequence also did not associate with any AP complex member, only sharing 364/470 (77%) identical sites with the consensus sequences.

Previously, the remaining two sequences associated with AP COI consensus, however with Cytb comparisons AP_yaw_01 closely associated with AK (100% identical to AK consensus) and AP_wip_05 did not associate with any Cytb consensus sharing 395/470 (84.0%) sites with the consensus sequences. Similar steps were taken to ensure the two un-associated sequences were of Anopheles origin. BLAST search top hits revealed high similarity to other Anopheles species.

Cytb sequencing was completed for 4 other individuals that were COI outliers (AF4_mad_06, AF4_mad_07, AP_dre_03, and AP_fin_01). Cytb analysis of these sequences demonstrated association with their previously identified species and was not congruent with COI consensus association observations.
e. Differentiation of the AP complex: How does the differentiation of the AP complex compare with other cryptic Anopheles groups?

Arlequin v3.5 (Excoffier and Lischer 2010) was used to assess differentiation of species within the AP complex, comparing the diversity between species. COI haplotype data generated using DnaSP was then utilized in Arlequin to calculate pairwise $F_{st}$ values, measuring the genetic differentiation between species. Haplotype data was created using an alignment of all AP complex members sequenced (n=97). $F_{st}$ values were generated as a measure of genetic differentiation. $F_{st}$ values range from 0-1 (no genetic divergence to high levels of divergence) (Hartl and Clark 1997).

AP complex $F_{st}$ values for inter-species comparisons are greater than 0.52 indicating genetic divergence exists in these populations (Table 5.7). The lowest values were found when comparing AFs.s. and AH (0.52885) or AF4 (0.60178), while the highest $F_{st}$ values were observed when comparing AK with AP (0.78586) and AH (0.79096). As a comparison, differentiation between the AP complex and an Asian malaria vector, An. dirus, was also assessed using Arlequin through $F_{st}$ analysis. The An. dirus haplotypes included in these analyses were compiled through a similar Anopheles COI survey by Walton and colleagues.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>AP</th>
<th>AFs.s.</th>
<th>AH</th>
<th>AF4</th>
<th>AK</th>
<th>An. dirus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFs.s.</td>
<td>0.73074</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AH</td>
<td>0.72709</td>
<td>0.52885</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF4</td>
<td>0.69215</td>
<td>0.60178</td>
<td>0.68282</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK</td>
<td>0.78586</td>
<td>0.73964</td>
<td>0.79096</td>
<td>0.75733</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>An. dirus*</td>
<td>0.85712</td>
<td>0.82835</td>
<td>0.84796</td>
<td>0.81391</td>
<td>0.87754</td>
<td>-</td>
</tr>
</tbody>
</table>

* Walton et al. (2000)
When Walton and colleagues calculated $F_{st}$ values for three species within the Dirus group (An. dirus A, C and D) they observed low differentiation between An. dirus A and D ($F_{st} = 0.048$), however comparing An. dirus A and D with An. dirus C revealed $F_{st}$ values $> 0.680$ indicating high levels of differentiation. Furthermore Walton and colleagues observed two distinct groups of An. dirus C and when comparing these two groups $F_{st}$ values were $> 0.421$ suggesting that these were two highly differentiated groups within the same species (Walton et al. 2000). Interestingly, genetic crosses between An. dirus D and both An. dirus A and C were found to be sterile. But genetic crosses between An. dirus A and C were found to produce fertile offspring (Morgan et al. 2010).

Haplotypes shared among more than one species were observed for this An. dirus survey, similar to the observations discussed previously in this chapter with the AP complex. There were not any shared haplotypes between Dirus and AP complex members. And not surprisingly, comparisons of the PNG collections with the Dirus haplotypes yield an $F_{st}$ of 0.62139; $F_{st}$ values are $> 0.81$ when comparing Dirus individually with each AP complex member.

f. Intra-species mismatch distribution predictive of phylogenetic signal

Computing the mismatch distributions, or the frequency of segregating sites within each species from pairwise comparisons, has been previously used to understand population characteristics in other Anopheles groups (Besansky et al. 1997, Walton et al. 2000). This approach was first described as a characterization tool in investigations of mitochondria (Slatkin and Hudson 1991) specifically within
human populations (Rogers and Harpending 1992). A uni-modal wave distribution of the pairwise mismatches is thought to be indicative of population expansion or the result of a population bottleneck. In contrast a ‘ragged’ distribution of pairwise mismatches is indicative of a stable population where lineage loss has caused ‘gaps’ in the distribution of sequence polymorphisms (Rogers and Harpending 1992). Arlequin v3.5 and DnaSP v5 calculate mismatch distribution and the ‘raggedness’

Figure 5.4 Mismatch distributions. Frequency distributions of observed pairwise nucleotide differences in the 1270bp of COI mtDNA sequence among intra-species comparisons of An. punctulatus (n=32), An. koliensis (n=19), An. farauti s.s. (n=17), An. hinesorum (n=13), and An. farauti 4 (n=16). Solid bars indicate observations of comparisons where as the black line represents the simulated observations as calculated by Arlequin v3.5 assuming population growth. All H_{ri} (raggedness) values had non-significant p-values indicating that the population expansion model cannot be rejected for the data sampled.
index (Harpending’s raggedness index \(H_{ri}\)). \(H_{ri}\) range values have not been clearly defined in the literature however can be described as the estimate of high-frequency variation within the sampled data. As the value of \(H_{ri}\) approaches 1, more ‘raggedness’ can be observed and more variation exists within the sequences. Mid-range \(H_{ri}\) values (e.g. 0.4) indicate a stationary population; \(H_{ri}\) values closer to 0 (e.g. 0.0004) indicate populations that have experienced (sudden) expansion. Because \(H_{ri}\) values are difficult to interpret, p-values for \(H_{ri}\) are a helpful statistic allowing acceptance (non-significant p-value) or rejection (significant p-value) of the hypothesis that the population has experienced expansion (Harpending et al. 1993).

In order to investigate population level characteristics of each of the five AP species in this dissertation, observed and expected frequencies of mismatches are illustrated for each species in Figure 5.4.

The most striking observation in Figure 5.4 is that of the differences between AK and the remaining 4 species. AK has a uni-modal distribution with an observed mean mismatch frequency of 16.77. Whereas the other four distributions appear to be more ragged. AP and AF4 are characterized by a single peak (max between 5 and 8 segregating sites) and then two smaller peaks at \(>\sim 90\) segregating sites. AFs.s. and AH both appear to have 3 peaks within the first 20 segregating sites. In AH these peaks are then followed by a peak at \(\sim 40\) and \(\sim 90\), where as AFs.s. appears to have two peaks at 80 and 90 respectively.

Simulated frequencies of mismatches were calculated by Arlequin to fit the data to a model of sudden population expansion and displayed as the black trend
line on each graph. This trend line seems to track well with the AK observed frequencies, and the first peak of AP and AF4. The later peaks in AP and AF4 do not fall under the simulation trend line. AFs.s. and AH are not tracked well by the simulation. This is due to the multiple peaks observed between 0-30 segregating sites. It is not possible to calculate the chi-square of the observed vs. simulated comparison, as many fields for observed and simulated contain zeros. However, $H_{ri}$ p-values for all five AP sibling species were non-significant suggesting that the population expansion hypothesis cannot be rejected.

The uni-modal curve from the AK analysis supports the observations made in Figure 5.1b where all 16 individuals seem to be very similar with no apparent outliers. The multiple peaks visualized between 0-30 of the AFs.s. and AH distributions supports the observation made in Figures 5.1A and 5.2, where outlier sequences are observed to be different from the majority of the remaining sequences. Two distinct populations were observed in AFs.s. (one population composed of 4 individuals from Manus island Province) along with 2 outlier sequences that were very different from the group. Two distinct populations were also found in AH in addition to having 2 outlier sequences in the alignment. These distinct populations were still more closely related to each other than to other species. The distant peaks in the AFs.s. and AH distributions can be attributed to the outliers identified in Figure 5.1A, 5.2 and 5.3. AP and AF4 appear to have a single peak between 0-30, which illustrates their appearance as a core set of very closely related sequences (Table 5.3). The outliers, previously identified and described (Figure 5.3), contribute to the distant peaks around 90-120 segregating sites.
When considering the uni-modal curve observed in AK and the first peaks of AP and AF4, AK peaks later (~17 segregating sites) than that observed for the other two (AP ~4 and AF4 ~7), suggesting that within AK a greater number of sequence comparisons harbor more differences than comparisons within the closely related AP and AF4 representatives. This result can be attributed to the greater diversity within AK (Table 5.5) and will be observable in the next chapter where branch lengths connecting species within the AK clade are longer than those connecting species within the AP and AF4 clades. The distant peaks within AP, AFs.s., AH, and AF4 mismatch distributions will also influence the phylogenetic analysis of COI. The representatives within each species that are contributing to these greater mismatch distributions will appear to associate with different species clades within the mtDNA comparisons.

D. Discussion

As a complementary analysis of nuclear sequences explored previously in this dissertation, this chapter focused on mitochondrial gene analysis. MtDNA has a high rate of evolution and is clonally inherited through maternal transfer. As a result analysis of mtDNA is thought to be useful for studying relationships of closely related or currently evolving species (Avise 1987, Hartl and Clark 1997). The entire mtDNA has been sequenced for 4 Anopheles species, two of which are in the same subgenus (Cellia) as members of the AP complex: An. gambiae and An. funestus. These genomes provided helpful references for amplification of mtDNA genes from the Anopheles punctulatus complex. Consensus sequences from each of the five AP
complex members were compared with the COI gene of these four species and demonstrated ~89% identity—equivalent to the pairwise similarities observed when comparisons between An. gambiae, An. funestus, An. darling, An. quadrimaculatus, (Table 5.2) and An. dirus are made.

To evaluate relationships among the AP complex members, portions of the Cytochrome Oxidase I (COI) and the Cytochrome B (Cytb) genes were analyzed for 97 and 37 individuals, respectively, from 21 villages in 7 PNG provinces (Table 5.1). First visual observations of alignments from these sequences illustrated that some representatives within each species appeared to be distinctly different from their own species. In most cases these differences were explained as the outlier sequences were more similar to other AP sibling species (Figure 5.1, 5.2 and 5.3). These observations were made more than once and for more than one species suggesting they were not laboratory or sequencing error. These observations could be evidence of gene flow (mitochondrial introgression) between these species or ancestral remnants of incomplete lineage sorting. In contrast since multiple individuals did not share the same discordant patterns and because this study is the first to report on COI and Cytb in a subset of AP sibling species, it is unclear that any definite evolutionary conclusions can be drawn from these observations.

Haplotypes were inferred for both genes, and haploid-based heterozygosity (H) values were high (H> 0.83) for both COI and Cytb in all species analyzed because each species harbored almost as many haplotypes as individual sequences analyzed (Tables 5.5 and 5.6). This suggested that a great deal of diversity (within these gene
regions) exists within members of the same species. Descriptions of inter-species pairwise differentiation ($F_{st}$) of the COI sequences revealed high genetic differentiation ($F_{st} > 0.52$) between members of the AP complex. Suggesting that these species are distinctly different, at least in the case of the COI sequence, further supporting a multiple species hypothesis for the five AP complex members. Additionally, the $F_{st}$ observations and interspecies sharing of haplotypes are consistent with observations made between other cryptic Anopheles species. $F_{st}$ values comparing An. dirus C with An. dirus A and D were similar to those generated in inter-species comparisons of AP complex members (Table 5.7). Within the AP complex, however, we do not see any two species comparisons that yield $F_{st}$ values like that reported by Walton et al. for An. dirus A and D ($F_{st} =0.048$) (Walton and Handley 2000). Besansky and colleagues have observed low $F_{st}$ values ($F_{st} =0.093$) when comparing mtDNA (ND5 gene) from two species of the An. gambiae complex (An. gambiae s.s. and An. arabiensis) collected from Senegal, Kenya, and South Africa. They concluded these low values were suggestive of gene flow between these species populations (Besansky et al. 1997).

Within the 97 individual mosquitoes sequenced for COI, a sub-population of AFs.s. from Manus Province (n=4), and island ~340 km off the northern coast of PNG was included. These 4 sequences were >99.2% similar to one another, however each contributed a unique haplotype to the total. Genetic differentiation was calculated comparing this population with the rest of the AFs.s. representatives (n= 13, from various mainland collections: Table 5.1). An $F_{st}$ of 0.41374 revealed a high level of differentiation between these two populations that is likely to be due to
the geographic isolation of the Manus samples. Mismatch distribution of AFs.s. samples suggested two distinct populations within the species representatives sampled that is directly related to the comparisons between the Manus and mainland samples (Figure 5.4).

Outlier sequences, which were more similar or shared haplotypes with other species, could possibly indicate evidence for gene flow between these species. However, studies completed in the early 1970s (between AFs.s. [East New Britain Province], AH [Queensland, Australia] and AK [National Capital District and Madang Province]) and 1980s (between AFs.s., AH and An. torresiensis [formerly An. farauti 3], all from Queensland), prior to genetic species identification methods, did not produce viable offspring from cross-mating (Bryan 1973a, Mahon and Miethke 1982). Cross-mating of other species within the complex has not been reported in the literature. Therefore, to verify these inter-species shared haplotypes, a larger sampling of these species is needed to achieve multiple observations.

Using the multiplex DNA probe approach, similar to that designed and described for other genes in Chapters 3 and 4, could be a potential method of monitoring for mitochondrial introgression in a large sample sets. It would be important to understand if introgression is occurring as it could pose a potential threat to integrated vector management strategies if insecticide resistance mutations, or other mutations increasing fitness in the face of vector management pressure, could be passed between species.
Data generated for this chapter allowed for a complementary genetic comparison of the AP complex using mtDNA and allowed for the calculation of genetic diversity and species differentiation. The data generated in this chapter is the first representation of mtDNA, specifically COI, among a diverse sampling of multiple individuals within five AP complex species across Papua New Guinea. The species differentiation data supports the multiple species hypothesis also supported by the two nuclear sequences in the previous chapters.
VI. CHAPTER 6

Using Phylogenetic Analysis to Assess Relationships Among *An. punctulatus* species complex Members
A. Introduction: A Phylogenetic Approach to Understanding Inter-An. punctulatus Group Relationships

A well-constructed phylogeny can be used to infer the evolutionary relationships for a group of taxa. This phylogeny can be used to visualize shared traits and may be useful in detecting phylogeographical patterns (Avise 2000). A phylogeny can also be used to make predictions regarding characteristics of one species based upon relationships to others species within the structure of a tree.

Even though Anopheles mosquitoes are of medical importance, because many species are vectors of malaria parasites and other deadly or debilitating disease agents, their evolutionary relationships remain largely ambiguous. In the most recent review of the Genus Anopheles, 485 species have been described (Harbach 2004), with at least a quarter of these belonging to a cryptic species complex/group.

![Figure 1.2 Phylogenetic relationships of the Culicidae Family with emphasis on subgenera in the Anopheles genus adapted from Besansky (1997) and Krzywinska (2001 & 2003). Estimated lineage splits are indicated at the following nodes: A. ~145-200 Mya (Krzywinska, 2006), 190 Mya (Moreno, 2010), B. 79 Mya (Moreno, 2010), C. 58 Mya (Moreno, 2010) to ~90-106 Mya (Krzywinska, 2006).]

Most species within the Anopheles Genus belong to the subgenera Cellia, Anopheles or Nyssorhynchus (Figure 1.2 from Chapter 1 has been included here for convenience) and each of these subgenera are further divided into
Series. In Papua New Guinea members of the *An. punctulatus* (AP) species complex are thought to be the main vectors of malaria and filarial parasites. A cryptic species group, the AP complex, belongs to the *Cellia* subgenus *Neomyzomyia* Series (Harbach 2004). *Neomyzomyia*, considered Old World species, are found throughout parts of Europe, Africa, Asia and the Pacific Islands are, but are absent from the Americas (Krzywinski and Besansky 2003, Harbach 2004, WRBU 2010).

Efforts to resolve the phylogeny of the *An. punctulatus* complex have encountered difficulties. Identification by morphological characteristics has described three species: *An. punctulatus* sensu lato (s.l.), *An. koliensis*, and *An. farauti* s.l., and because distinguishing features can be polymorphic in coloration, species are often misclassified (Cooper et al. 2002, Benet et al. 2004b, Henry-Halldin et al. 2011). However, molecular and cross-mating studies suggest this species complex is composed of up to 12 distinct species (Bryan 1973b, Mahon and Miethke 1982, Mahon 1983, Foley et al. 1993, Foley et al. 1995, Cooper et al. 2000, Cooper et al. 2009a). Three previous attempts to reconstruct species relationships have been made that each focused on a different gene (discussed in Chapter 1); two nuclear sequences (18SrDNA and ITS2) and a mitochondrial sequence (*COII*). Figure 1.10 (from Chapter 1 has been included here for convenience) compares the phylogenies produced from these three manuscripts. In general, bootstrap support for most of the nodes in all three trees are below 85. The placement of *An. koliensis*, as well as the relationships among clades, is different for each phylogeny.
For each of these phylogenetic comparisons only a single representative sequence for each species was used (Foley et al. 1998, Beebe et al. 1999b, Beebe et al. 2000a). As demonstrated in Chapter 5, occasionally, sequenced mitochondrial genes from AP complex members have been shown to more closely associate with a species different from their own identification. Therefore the use of only one representative per species, especially in the case of mtDNA, could potentially interfere, if misidentified, with the assembly of an accurate *An. punctulatus* species complex phylogeny.
Moreover, relying on a single *gene* to describe an accurate species phylogeny can be problematic as genes and species are evolving at the same time. While a tree created from a single gene analysis (gene tree) may portray an accurate representation of the evolution of that specific gene, it may not represent the evolutionary path of the species (species tree) (Degnan and Salter 2005, Degnan and Rosenberg 2006). This is potentially an explanation for the different tree topologies resulting from the previously published phylogenetic hypotheses (Figure 1.10). Involving multiple genes in a single analysis is an approach that is currently being used to mediate the gene tree vs. species tree issue (Pevsner 2003).

Aim 2 of this dissertation seeks to determine if relationships within the AP complex can be verified using phylogenetic techniques. In this chapter I hypothesize that an improved phylogeny of the AP group can be achieved using multiple DNA sequences. I will test this by using DNA sequence data generated in Chapters 3, 4 and 5. As previously described, these DNA sequences are on different chromosomes and/or are localized to nuclear or mitochondrial DNA and are of different and function. By analyzing each gene individually it is possible to understand the evolution of these specific sequences and compare/contrast these gene tree topologies to one another. Combining multiple different DNA sequences for analysis can potentially produce an improved phylogeny, as the evolution of each individual gene will be incorporated to attempt a better estimate of a species tree.

Additionally, I hypothesize that phylogeographical patterns will immerge within the *An. farauti* s.s. collection populations. This species has been sampled on the mainland—collections primarily within 10km of the coast and on Manus
Island—340 km north of the PNG mainland. Finally, by utilizing previously published understanding of insect mitochondrial DNA evolution and available mtDNA sequence, I want to test if an approximate time of species divergence within the AP complex can be inferred and compared to other mosquito lineages.

B. Methods

I have chosen to focus on two general methods in this dissertation, neighbor-joining (NJ) and maximum likelihood (ML). These methods are well described in the literature and the trees provide modest, uncomplicated assumptions (Xiong 2006, Pevsner 2003), which is helpful since not much is known about the AP complex. Previous phylogenetic reconstructions (Figure 1.10) of the AP sibling species used variations of parsimony analysis and were not successful in reconstructing a well-supported phylogeny. Because of this, I wanted to test if by employing different methods (NJ and ML) an improved phylogenetic tree could be created for the five AP sibling species analyzed in this dissertation.

NJ is a distance-based method that is described in greater detail in Chapter 2 Section H. This method uses a distance matrix to create a tree in a step-wise manner, comparing each sequence to the rest of the group, one at a time. Because this method is not computationally taxing it is considered to be useful when analyzing large data sets. However, because this method is thought to be rudimentary, the most accurate representation of evolutionary relationships may not be captured as NJ takes only one path to create the relationship tree (Xiong
Because of this limitation NJ is not the only method that is used in this dissertation.

The program MEGA 5 was used for NJ analysis of ITS2, VGSC, COI and Cytb sequence alignments (Tamura et al. 2011). In order to decide if NJ is an appropriate analysis tool for the data set, Nei and Kumar, designers of the MEGA phylogenetic analysis program, suggest that the average pairwise Jukes-Cantor distance should be less than 1.0 (Nei and Kumar 2000, Hall 2008). Jukes-Cantor distance for all 4 DNA sequences was calculated to be less than 0.115 (ITS2=0.115, VGSC=0.070, COI=0.067, Cytb=0.058), suggesting that NJ analysis was appropriate for these data.

Deciding on a model of substitution is necessary before any phylogenetic analysis can proceed. For NJ analysis two different models of substitution were chosen to represent a simple (Jukes-Cantor) and more sophisticated (Maximum Composite Likelihood [MCL] a version of the Tamura-Nei model) and have been discussed in Chapter 2 Section H.

jModelTest (Posada 2008, 2009), a program that will determine the best-fit nucleotide substitution model for the data was used to assess the best-fit model for use in ML tree building. The best-fit model was determined by the AIC (Akaike Information Criterion) calculated by jModelTest. AIC compares the results of all models at the same time to find the model that, when implemented, causes the least amount of information loss while assessing penalties for ‘over-parameterization’ (Anderson and Burnham 2002, Sullivan and Joyce 2005, Posada 2009)—if two models explain the data equally well, the simpler of the two will be preferred. In addition to choosing a substitution model, statistical evaluation of the
reproducibility of trees created in this dissertation was carried out by the bootstrapping method (Felsenstein 1985) that is also further described in Chapter 2 Section H. Bootstrap values range from 1-100, and can be considered as the percent or proportion of replicate trees that harbor the specific node given in the consensus tree, therefore bootstrap values above 80 will be displayed in this dissertation, with high support considered to be >90.

RAxML (Randomized Axelerated Maximum Likelihood) v7.2.7 (Stamatakis et al. 2002, Stamatakis 2006) was used to approximate a species tree using maximum likelihood analysis. RAxML will accept a partitioned dataset where multiple genes are included for the same individual. Each gene is analyzed separately, and then analyses are combined to produce a consensus tree. This method is designed to handle large numbers of sequences and produce a ML tree in a more efficient manner than standard ML approaches. Because ITS2, VGSC and COI were sequenced most often for the 111 individual mosquitoes included in this dissertation, the alignments of these DNA sequences were concatenated into a continuous partitioned dataset using Geneious 5.3. Data was submitted along with identification of gene partition locations to RAxML running on the Cipres Portal HPC (Miller 2009).

For estimating divergence times within the AP complex a Bayesian approach using BEAST (Bayesian Evolutionary Analysis by Sampling Trees) v.1.6.1 (Drummond and Rambaut 2007) used previously published molecular divergence times of Drosophila, Aedes, and An. gambiae and An. funestus to create a chronogram (using COI) including the AP complex members (Moreno et al. 2010). BEAST
reconstructs phylogenies without relying on a single tree topology to begin comparisons. BEAST utilizes Markov chain Monte Carlo simulation (MCMC) to continuously sample topology outcomes so that the final tree has a topology that is weighted proportionally to its clade’s posterior probability (Drummond and Rambaut 2007). BEAST companion programs were used to create the input file (BEAUti v1.5.3), visualize the quality of the MCMC runs (Tracer v1.5), and calculate the most credible tree (TreeAnnotator v1.5.3). Statistical evaluation of tree node confidence in BEAST analysis is evaluated using posterior probabilities.

Four DNA sequences generated and described in Chapters 3, 4 and 5 were used in this chapter for phylogenetic analysis. All sequence alignments were generated in Chapters 3, 4, and 5 using ClustalW (Larkin 2007). Nuclear sequences included a transcribed region of rDNA, the Internal Transcribed Spacer Unit 2 (ITS2) that separates 5.8 and 28S rDNA coding regions. This spacer region is thought to evolve at a faster rate than the flanking coding regions and because of this, it is thought to be a good candidate for analyzing phylogenetic relationships between the closely related members of a species complex. ITS2 is also the main molecular target for species discrimination by PCR based methods (Beebe and Saul 1995, Henry-Halldin et al. 2011). My analysis includes a portion of the Voltage Gated Sodium Channel gene (VGSC), a nuclear coding sequence. This gene sequence was previously described in detail (Chapter 4) and found to harbor numerous nonsynonymous exon SNPs and diverse intronic regions.

Mitochondrial sequences included Cytochrome Oxidase I (COI) and Cytochrome B (Cytb) and were described in Chapter 5. Mitochondrial genes are
thought to provide a potentially different viewpoint of species evolution since they are maternally inherited and lack recombination (Avise 1987, Hartl and Clark 1997). Mitochondrial genes are also frequently used in phylogenetic analysis because they are rapidly evolving and may provide a more accurate depiction of species relationships, especially among closely related species (Moore 1995, 1997). However, the accuracy of mtDNA to define the 'best' species relationships has been refuted (Hoelzer 1997). Therefore, no 'gold standard' gene or DNA sequence exists for reconstructing phylogenetic relationships, which is why DNA sequences of different origin and function were chosen for phylogenetic analysis in this chapter.

The sequenced data used contained many individual mosquitoes representing 5 species within the AP group (n=111; *An. punctulatus* [AP]=34, *An. farauti* s.s. [AFs.s]=18, *An. hinesorum* [AH, also known as *An. farauti* 2 or AF2]=22, *An. farauti* 4 [AF4]=16, *An. koliensis* [AK]=21) that were represented by at least one of the four DNA sequences, with over 74% of the individuals represented in 3 or 4 of the sequence alignments. These five species represent the most common and widespread AP species throughout the country (AP, AK, AFs.s., AH) and specifically (AF4) within our study locations. These mosquitoes are also considered to be the most important vectors of human malaria and filariasis in PNG.

<table>
<thead>
<tr>
<th>DNA Alignments</th>
<th>Parsimony-Conserved Sites</th>
<th>Table 6.1 Details of DNA Sequence Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>Aligned Length*</td>
<td>Informative Sites</td>
</tr>
<tr>
<td>ITS2</td>
<td>897</td>
<td>229</td>
</tr>
<tr>
<td>VSCG</td>
<td>1448</td>
<td>406</td>
</tr>
<tr>
<td>COI</td>
<td>1270</td>
<td>250</td>
</tr>
<tr>
<td>Cytb</td>
<td>470</td>
<td>77</td>
</tr>
</tbody>
</table>

* nucleotides

As an outgroup, *An. longirostris*, from PNG collection sites, was included in sequencing efforts (Chapters 3-5). *An. longirostris* (AL) is not a
member of the AP complex, however it is taxonomically part of the same Series (Neomyzomyia) of the Cellia Subgenus and is also found in sympatry with members of the AP complex. For the mitochondrial genes analyzed the An. gambiae mitochondrial genome sequence was used, in addition to AL, as an outgroup. An. gambiae was not used as an outgroup for VGSC or ITS2 because the same regions for the VGSC were not available for An. gambiae in Genbank and the ITS2 sequence was notably different: ~440bp in length compared to ~700 bp for AP complex ITS2 sequence length. DNA sequence alignment information (Table 6.1) was calculated using MEGA5 (Tamura et al. 2011). All sequences have been submitted to Genbank (accession information Table 6.2).

<table>
<thead>
<tr>
<th>DNA Alignments</th>
<th>Genbank Accession Nos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITS2</td>
<td>HM584365-HM584454</td>
</tr>
<tr>
<td>VGSC</td>
<td>HQ173903-HQ173993</td>
</tr>
<tr>
<td>COI</td>
<td>HQ840792-HQ840889</td>
</tr>
<tr>
<td>Cytb</td>
<td>HQ840890-HQ840928</td>
</tr>
</tbody>
</table>

### C. Results

#### 1. Neighbor-Joining analysis

Based on sequence alignments for each of the four DNA sequences described in Chapters 3, 4 and 5, NJ phylogenetic trees were created using MEGA5 (Tamura et al. 2011) applying Jukes-Cantor and Maximum Composite Likelihood (a version of the Tamura-Nei model that has been discussed in Chapter 2 Section H) substitution models with 1000 bootstrap replicates. The results from using these two substitution models were almost identical (same clade topology and bootstrap values varied by one or two points at nodes); therefore only one tree will be shown for each sequence (Figure 6.1). Phylogenetic trees are shown in Figure 6.1 where
individual sequences were assigned an identifier code based upon their prior species identifier (via ITS2 LDR-FMA, VGSC LDR-FMA and/or ITS2 sequence: See Chapters 3 and 4), the location they were collected and a number to distinguish multiple representatives of the same species from the same location. For example the third AP representative from Dreikikir (Table 2.1) was given an identifier of AP_dre_03. All individuals were given the same identifier for each gene sequenced, therefore the ‘AP_dre_03’ sequence in the ITS2 tree is from the same individual as the ‘AP_dre_03’ in the VGSC, COI an Cytb trees.

a. Nuclear sequence results

The ITS2 tree contained 89 (out of the 111 individuals analyzed in this dissertation) members of the AP species complex (AP=27, AFs.s.=15, AH=12, AF4=16, AK=19) and the VGSC tree contained 90 AP complex individuals (AP=31, AFs.s.=14, AH=13, AF4=12, AK=20). A majority of these individuals (n=75) were represented in both nuclear sequence (ITS2 and VGSC) alignments. Three AL representatives were sequenced for ITS2 and VGSC and served as outgroup comparisons in both trees. Both nuclear sequences demonstrated well-supported clades for each of the 5 species included in the analysis with internal nodes having >95 bootstrap value. For the ITS2 phylogeny AFs.s. and AH are connected by a node with a bootstrap value of 100, this same node has 99 bootstrap support with the VGSC phylogeny, suggesting strong support for a shared common ancestor between AFs.s. and AH. In the ITS2 tree the relationship between AFs.s./AH and AK clades is supported by a node with a 99 bootstrap value. This relationship was also observed in the VGSC tree but did not have strong support (74 bootstrap). Interestingly, this
Figure 6.1 **Neighbor-joining phylogenetic trees** constructed from two mitochondrial and two nuclear genes: Internal Transcribed Spacer 2 (ITS2) (alignment length 897 bp), Voltage Gated Sodium Channel (VGSC) (1448 bp), Cytochrome Oxidase I (COI) (1270 bp), Cytochrome B (Cytb) (470bp). Bootstrap values >80 are shown. Clades are colored according to species: AF.s.s. (‘AF1’)=Red, AH (‘AF2’)=Orange, AF4=Purple, AK=Green, AP=Blue. Sequences belonging to a discordant clade are indicated with asterisks (colored according to their species identification [via ITS2 sequence and/or ITS2 & VGSC LDR-PMA]).
relationship has not observed in Beebe et al.’s ITS2 phylogenetic reconstruction (Beebe et al. 1999a) (Figure 1.10). The node connecting AP and AF4 in the ITS2 tree was also not well supported (75 bootstrap), this node was also present in Beebe’s tree and lacked significant support.

Relationship topology within the VGSC tree is similar to that of the ITS2 tree. An exception to this general summary finds that AF4 and AP are not found to share a common node. Bootstrap support for this node in the ITS2 tree is low. Support for the node connecting AF4 with the AFs.s./AH/AK clades is also low (52 bootstrap). With both trees it is clear that the identifications given to species are consistent with their placement within the clades of the phylogeny. For example AFs.s. representatives are not found in the AP clade. It is also clear that the AL outgroups are distinctly different from the AP complex members.

b. Mitochondrial sequence results

The COI tree contained 97 members of the AP species complex (AP=32, AFs.s.=17, AH=13, AF4=16, AK=19). The Cytb tree contained the least individuals with 37 AP complex representatives (AP=6, AFs.s.=4, AH=18, AF4=6, AK=3). At the time this gene was being amplified, stocks of DNA for the 111 individuals used for this dissertation had been depleted for other purposes in the laboratory. At least one AL representative was used as an outgroup for both genes. COI and Cytb sequences were extracted from the published An. gambiae mitochondrial genome (Beard et al. 1993) in Genbank (NC_002084) and used as an outgroup for these mtDNA-based analyses. Of the 103 individual mosquitoes represented in the
mtDNA trees, 31 were represented in both COI and Cytb trees. Figure 6.2 is a graphical demonstration showing that 98 individual mosquitoes in this dissertation were represented by at least one mtDNA and one nuclear DNA sequence (ITS2 or VGSC) in the phylogenetic trees (Figure 6.2).

In general mtDNA analysis supported the existence of five species (COI tree 99 bootstrap, Cytb tree bootstrap>92) (Figure 6.1). The COI tree, similar to the ITS2 and VGSC trees, has strong support (99 bootstrap) for AFs.s. and AH as sister taxa. This node is also present in the Cytb tree but with much less support (69 bootstrap). The placement of AP, AK, and AF4 in the COI tree has an organization not observed in the ITS2 or VGSC trees: AF4 is sister to the AFs.s./AH clades and, AK and AP are hypothesized to share a common node. Bootstrap support here is also low (46 & 48 respectively) therefore ancestry cannot be assumed by these nodes. Cytb has a topology similar to VGSC but bootstrap values supporting these nodes range from 22-40.

Upon closer inspection of the clades within the COI tree, AFs.s., AH, AF4 and AP individuals were found in clades outside their species identification. These
disparate sequences are denoted by colored asterisks in Figure 6.1. Single AFs.s. individuals are found in the AF4 and AK clades, one AH individual is found in the AP clade, two AF4 individuals are found in the AFs.s. clade, and three AP individuals are found outside the AP clade: two in the AF4 clade and one associating with the AL outgroup. These disparate sequences are the same individuals that were previously labeled ‘outliers’ in the COI diversity analysis of Chapter 5 (Figure 5.3), contributing to multiple peaks in the mismatch distributions of each species (Figure 5.4). Therefore, their behavior in these phylogenetic analyses is not of a surprise. The only previously identified ‘outlier’ that did not appear to be disparate in the COI tree is AF2_wip_01 (also known as AH_wip_01). AF2_wip_01 was found in close association with the AH clade, however this sequence did not group with the same close association as other members of the AH clade. Though bootstrap support for the AH clade is high (99 bootstrap) the node connecting AF2_wip_01 and the rest of the clade has low support (60 bootstrap).

A closer look at the Cytb tree reveals similar patterns of disparate sequences, even though it contains just over one third of the number of individuals included in the COI tree. AF2_nal_01, which associated with the AP clade in the COI tree, also associates with the AP clade in the Cytb tree. The outlier, AF2_wip_02, described in the previous paragraph, appears to be very distant from the rest of the AP complex however branch support is low (66 bootstrap). Two AP individuals that were represented in the AP clade of the COI tree appear disparate in the Cytb tree associating with the AK clade (AP_yaw_01) and AL clade (AP_wip_05). All of these
disparate individuals were previously labeled as ‘outliers’ in the Cytb diversity analysis of Chapter 5.

Chapter 5 indicated that all disparate (or ‘outlier’) sequences were inspected for evidence of being nuclear pseudogenes and run through BLAST (http://blast.ncbi.nlm.nih.gov/) to check that they were of Anopheles origin. All disparate sequences (with the exception of COI sequence AP_fin_01 and Cytb AF2_wip_01) shared high homology (most >99% pairwise identity) with a clade in the AP complex.

c. **Phylogeographical observations within of An. farauti s.s. and An. hinesorum**

In aim 2 it was hypothesized that phylogeographical patterns would immerge between populations separated by geographical barriers, specifically within An. farauti s.s. due to the distance and body of water separating collection sites on the mainland and on Manus island. Phylogeographical signal can be observed within the AFs.s. clade where the individuals from Lorengau on Manus island (n=4; AF1_lor_01, 06, 08, 10; indicated in tree with red bar), located 340 km from the north coast of PNG, are separated from the rest of the clade composed of AFs.s. from mainland PNG, mostly from around Madang (Madang Province) or Godowa (Morobe Province) (Figure 6.3). This separation is supported by a high bootstrap value (97) at the connecting node. This separation was also observed in Chapter 5 with the mismatch distribution and $F_{st}$ (value of 0.4) calculations between mainland and Manus AFs.s. populations.
Interestingly, similar observations are found within the AH clade of the Cytb and VGSC trees. Eleven samples from Ramu (AF2_rmu; Morobe Province) along with two other individuals, one from Singoropa (AF2_sin_01; Western Highlands Province) and one from Nale (AF2_nal_02; East Sepik Province) appear to form a grouping separated from the three samples from Kuru (AF2_wip_02, 10, and 12; Western Province; indicated in tree with orange bar). These clades are supported by bootstraps of 92 and 99, respectively. Ramu (collected at an altitude of 444 m) and Nale (104 m) collection sites are both found north of the Central Ranges. Kuru (43 m), a collection made south of the Central Ranges is separated from Ramu and Nale by at least 450 km. The individual from Singoropa, located in the Western
Highlands Province, was captured at ~1300 m on the northern side of the Central Ranges (~300 km from Nale, ~150 km from Ramu and ~400 km from Kuru) (Figure 6.3). $F_{st}$ was calculated comparing populations north of the Central Ranges to those collected south of the Central Ranges and these populations were found to be distinctly different ($F_{st}=0.3238$).

d. Summary of Neighbor-Joining results

Generally, the NJ analysis performed on individual nuclear and mitochondrial sequences supported the multiple species hypothesis. Clades for AP, AFs.s., AH, AF4 and AK were well supported (bootstrap >92) within each gene tree. In contrast, ancestral relationships could not be derived from the phylogenies because basal relationships had very low support. The only relationship between clades that were observed consistently and with high support was the node connecting AFs.s. and AH, suggesting that these two sequences share a common ancestor.

Previously, the ITS2 DNA sequence had been used to reconstruct the AP species complex phylogeny through parsimony analysis (Beebe et al. 1999a) (Figure 1.10). To compare the ITS2 sequences generated in this dissertation previously reported sequences were obtained from Genbank (Accession nos. AF033213-AF033221) and added to the ITS2 sequence alignment used in this dissertation. NJ (assuming the MCL substitution model) phylogenetic reconstruction was performed as previously described by MEGA. All previously published ITS2 sequences belonging to AP, AFs.s., AH. AF4 and AK were found to integrate into the expected clades with the individuals sequenced in this dissertation (Figure 6.4). The addition of the 5 other AP species sequences (An. farauti 3, 5, 6, 7 and An. species near
*punctulatus* did not alter the topology of the ITS2 tree, nor did it improve the bootstrap support at ancestral nodes. The topology of this tree with the additional AP complex members resembles the phylogeny previously published by Beebe and colleagues in that AFs.s. and *An. farauti* 7 share a common ancestor. AH, *An. farauti* 5 and *An. farauti* 6 and *An. farauti* 4 and *An.* species near *punctulatus* appear to be closely related. Unique to the ITS2 tree is the highly supported (97 bootstrap) node connecting AFs.s./*An. farauti* 7 clade with AH/*An. farauti* 5/*An. farauti* 6 clade (Figure 6.4). This relationship is not observed in ITS2 phylogeny reported by Beebe et al. (Figure 1.10).
2. **Maximum Likelihood analysis**

As indicated previously results from the NJ analysis were unable to resolve a number of ancestral relationships within the AP complex. Because NJ analysis potentially lacks functional capacity to produce the most evolutionarily relevant tree, I applied a more complex tree optimization approach to the 4 sequence alignments analyzed in the previous section. Maximum Likelihood (ML) analysis was performed by MEGA using the best-fit substitution model (lowest AIC score computed by jModelTest: *COI, Cytb* and *VGSC=GTR+G* and *ITS2=HKY+G*) and 1000 bootstrap replicates (Figure 6.5).

a. **Nuclear sequence results**

ML analysis of nuclear sequences (*ITS2* and *VGSC*) produced the same topology as NJ analysis with respect to species clades. The connecting node of AF4 and AP had higher bootstrap support in the ITS2 ML tree when compared with the NJ tree for ITS2 (ML bootstrap 83 vs. NJ bootstrap 74), however bootstrap for the connecting node of the AFs.s./AH clade and the AK clade was slightly lower in the ML analysis (92 bootstrap vs. 99 bootstrap). Overall, ML analysis of nuclear sequences was consistent with NJ analysis.

b. **Mitochondrial sequence results**

Like NJ analysis, the older nodes were not highly supported; therefore low confidence is placed in hypothesized relationships between these clades. Similar to NJ analysis the ML tree found high bootstrap support (>98) is given for individual clades and the node connecting the AFs.s. and AH clades (93 bootstrap).
Figure 6.5 Maximum likelihood phylogenetic trees constructed from two mitochondrial and two nuclear genes: Internal Transcribed Spacer 2 (ITS2) (alignment length 897 bp), Voltage Gated Sodium Channel (VGSC) (1448 bp), Cytochrome Oxidase I (COI) (1270 bp), Cytochrome B (Cytb) (470bp). Bootstrap values >80 are shown. Clades are colored according to species: AFs.s. ('AF1')=Red, AH ('AF2')=Orange, AF4=Purple, AK=Green, AP=Blue. Sequences belonging to a discordant clade are indicated with asterisks (colored according to their species identification[via ITS2 sequence and/or ITS2 & VGSC LDR-FMA]).
ML analysis of the *Cytb* gene demonstrated strong support for monophyly of most species (bootstrap >90 for AK and AF4, >80 for AP, AFs.s. and divisions within AH) even though only one third of the 111 individuals were included in *Cytb* analysis (Figure 6.5). The same ‘outlier’ sequences were also visualized in the ML analysis as they are placed in the same clades as both mtDNA NJ analyses. Support for two sub-clades within the AH clade is given by bootstrap values of 84 (Ramu, Nale and Singropa) and 100 (Kuru). This phylogeographic pattern was observed in NJ analysis where these collection sites were described as separated by at least 300 km and in some cases separated by the Central Ranges (Figure 6.3).

3. **Gene Trees to Species Tree: RAxML simultaneous analysis of three genes to infer species tree**

In an attempt to approximate a single species tree, ITS2, *VGSC* and *COI* were analyzed using RAxML v7.2.7 and a single tree was produced. Figure 6.6A illustrates the output from RAxML using all 111 individuals. Through this analysis it became clear that the previously identified disparate individuals (outliers; Figures 6.1 and 6.5) were strongly influencing the phylogeny when analyzing all 111 individuals. The branching topology, especially with regards to AK and AFs.s., was very different from what had been observed in NJ and ML analysis with entire clades branching from one sequence (e.g. the entire AK clade is nested within the AFs.s. clade). Additionally nodes supporting AF4, AFs.s., and AH clades had low support, whereas in previous gene trees they had received high support. We cannot be certain of the origin for the rogue individuals observed in the mitochondrial gene analysis.
Therefore because these rogue individuals severely affected the topology of the tree they were removed (Thomson and Shaffer 2010) from the RAxML analysis and the pruned sequence alignment (n=102 individuals) was used for a second round (Figure 6.6B). This pruned analysis produced a better supported tree, free of rogue taxa, where each species clade was supported with a bootstrap of 100, with the
exception of AH (AF2), which was still monophyletic (84 bootstrap) and was divided into two distinct clades with bootstrap support of 99 and 85. This analysis placed AFs.s. and AH as sister clades with high support (100 bootstrap). The node that joined AFs.s./AH and AK was also highly supported (100 bootstrap). Both AP and AF4 clades were supported by 100 bootstrap values, however the node connecting these clades was not well supported (62 bootstrap). The node connecting AFs.s./AH/AK with AP/AF4 was highly supported (100 bootstrap).

This analysis revealed a strongly supported relationship between AFs.s./AH and AK. This relationship was previously observed in both ITS2 and VGSC NJ and ML analysis and was only well supported in the ITS2 trees, however it was not observed in previously published analyses. This approach to resolving the phylogeny of these five species, however, did not solidify AP and AF4’s placement.

4. Approaches to estimate divergence times

Divergence times between *Anopheles* species, *Aedes* (a vector of the human disease yellow fever and dengue viruses: member of the Culicidae Family, Subfamily Culicinae) and *Drosophila* (fruit fly) mitochondrial genomes have been published. Here I will use these times as calibration points and assume similar rates of evolution have occurred for sibling species of the AP complex. I will then infer divergence times for members of the AP complex using BEASTv1.6.1. For this analysis the pruned *COI* alignment (n=87) used in RAxML analysis was first aligned with the *COI* sequence from *Drosophila yakuba* (Genbank Accession X03240), *Aedes aegypti* (NC010241), *An. gambiae* (NC02084) and *An. funestus* (NC008070). This alignment was subjected to Bayesian analysis utilizing BEASTv1.6.1 on a web-based
High Performance Computing resource (BioHPC) through the Computational Biology Service Unit at Cornell University.

In the RAxML output, using the pruned data set, each species created distinct, well supported clades. This was also true for individual gene analysis using both NJ and ML approaches. BEAST has the capability of accepting trait values, including species-based data. This allows the program to assess a ‘species tree’ from DNA data that may contain several representatives of one species, like the data set in this dissertation. Therefore during the creation of the input file individuals were organized by specific species—this will also create a visually uncomplicated tree where each of the five species within the AP complex will be given one node, instead of all 98 sequences being displayed in the tree.

GTR+Gamma was used as the substitution model (chosen previously by jModelTest), a relaxed uncorrelated lognormal clock was used and the Yule tree prior was used as it has been suggested as a suitable speciation model for trees involving different species (Drummond et al. 2006, Drummond 2007). This approach has also been used to calculate divergence time between species of mosquitoes previously (Moreno et al. 2010). An MCMC chain of 30 million generations (sampled every 1000 generations) was indicated. The input file containing all prior information and sequence alignment was submitted to BioHPC. Following the run, FigTree v1.3.1 was used to visualize and annotate the maximum clade credibility tree (Figure 6.7).

Posterior probabilities range from 0-1, with 0 as low support for the placement of the node and 1 as high support for the node. Results of this analysis
show that supporting posteriors were high for all nodes in the tree except the node supporting the placement of AK (0.6459). Previously reported divergence times between *Drosophila* and *Anopheles* (~259.9 MYA) (Gaunt and Miles 2002), *Aedes/Culex* and *Anopheles* (~145-200 MYA) (Krzywinski et al. 2006, Moreno et al. 2010) were used to calibrate the tree. This calibration method estimates that AH diverged from the group ~6.7 MYA followed by AK (~3.8 MYA), AP (~2.23 MYA) and suggests that AFs.s. and AF4 are the youngest species resented in the AP sibling species analyzed, diverging ~1.5 MYA.

![Figure 6.7 BEAST Chronogram using Cytochrome Oxidase I (COI) (1270 bp) alignment (AP species n=87), Calibration for *Drosophila-Anopheles* split (~260 MYA) and *Aedes/Culex-Anopheles* split (~190 MYA) (Gaunt 2002, Krzywinski 2006, Moreno 2010) mostly based upon whole mitochondrial analysis. Branch support, in the form of posterior probability, is indicated at branch nodes.](image-url)
D. Discussion

In an attempt to reconstruct the evolutionary relationships of five *An. punctulatus* species complex members, four different methods of phylogenetic reconstruction were applied to DNA sequences from 111 recently collected individual mosquitoes collected throughout Papua New Guinea (Table 2.1, Figure 6.3). The rationale behind the DNA sequences chosen for analysis was described previously in Chapters 3, 4 and 5. The four DNA sequences were chosen to represent different origins (nuclear vs. mitochondrial) as well as different function. Previously, single-gene phylogenetic analysis of the AP complex had been attempted using three different genes (Figure 1.10). Each model produced a different topology with generally unsupported nodes. The purpose of this chapter and goal of this dissertation’s aim 2 was to utilize several different phylogenetic methods to reconstruct a well supported phylogenetic tree and to make inferences about divergence times between the species using previously published time calibration estimates.

1. Phylogeographic patterns

I hypothesized that because of geographical distance (including a body of water separating populations) phylogeographical patterns for known isolated populations (i.e. *An. farauti* s.s. from mainland vs. Manus island) could be detected. The phylogenetic analysis in this chapter demonstrated that *An. farauti* s.s. and surprisingly, *An. hinesorum* both demonstrated phylogeographical signal. The AFs.s. clade bifurcated into two groups, one containing only mainland mosquitoes and the second only mosquitoes collected on Manus island, within both NJ and ML analysis.
of COI with high bootstrap support. Phylogeographic observations within An. hinesorum were somewhat unexpected, and individuals tended to cluster within two groups: those collected north of the Central Ranges and those collected south of the Central Ranges. In retrospect this observation can be explained by the presence of the Central Range.

Phylogeographic patterns within AFs.s. collections suggested that AFs.s. sequences from Manus island (~340 km from the mainland Figure 6.3; sample identifiers AF1_lor_) harbor geographically distinct polymorphisms, especially within COI, which are not found in the mainland AFs.s. sampled. AF1_lor_01, 06, 08 and 10 are found to cluster in their own clade in both (NJ and ML) COI trees (Figures 6.1 and 6.5 Manus individuals labeled with red bar) with high support (97 and 99 bootstrap support respectively). This clustering was also observed in the multi-gene RAxML analysis but not highly supported. Reduced support observed in the RAxML analysis is likely because a distinct division did not exist within AFs.s. for the two nuclear genes (ITS2 and VGSC) analyzed.

More distinct phylogeographic signal was observed within the AH clade throughout most genes and analysis techniques. Figure 6.3 highlights the collection locations of AH (orange arrows), where samples from Singoropa, Ramu and Nale (all collected to the north of the Central Range) cluster together in the trees separate of samples collected from Kuru (collected near where the Fly River delta empties into the Gulf of Papua in the southern part of Western Province). An. hinesorum is generally a species found inland and in lower lying river valley habitats. Cooper and colleagues have documented An. hinesorum as the most abundant AP species south
of the Central Ranges. It is also important to note that AH is found north of the central ranges but in less abundance and in association with other species (Cooper et al. 2002).

We can hypothesize that the Central Range isposing a physical and climatic barrier, inhibiting gene flow between populations of *An. hinesorum* to the north and to the south of the steep mountain range. AP sibling species are not commonly found above altitudes of 1000 m, are rarely observed above 1500 m. The highest altitude where a member of the AP complex has been observed for *An. farauti 6*, thought to be adapted to the highlands, collected at 2000 m. The highest elevation an *An. hinesorum* has been collected is 1740 m (Cooper et al. 2002). The steep Central Range has many peaks exceeding >4000 m creating a cooler climate and therefore may form a physical boundary that *Anopheles* species adapted to low-lying warmer environments cannot traverse.

### 2. Phylogenetic relationships within the AP sibling species

Overall Neighbor-Joining and Maximum Likelihood analysis produced well-supported clades in all 4 DNA analyses (Figures 6.1 and 6.5). However each DNA sequence yielded a unique tree topology with respect to relationships between species. Nodes connecting species were not well supported by a high bootstrap value. The only species relationship that was consistent and well supported was the node between AFs.s. and AH found in all four DNA trees (both NJ and ML) with high support (ITS2, *VGSC*, *COI* >99 bootstrap). Within the ITS2 tree the AK clade and the AFs.s./AH clade were found to be sister clades with a high support (NJ 99 bootstrap, ML 92 bootstrap). This relationship was however not repeated (with support) in
other trees (NJ and ML). When analyzing ITS2, VGSC and COI in a concatenated data set, RAxML results suggest that AK is closely related to AFs.s. and AH. However RAxML analysis was also not able to place AP and AF4 in the species relationships with high support. NJ, ML and RAxML analysis was not successful in finding a well-supported relationship among the all 5 AP sibling species. This can possibly be attributed to ‘gene tree-species tree’ differences, where specific genes take different evolutionary paths that may not correspond directly to the evolution of the species (Degnan and Rosenberg 2006). Therefore individual gene trees only describe the evolution of the specific gene and possibly do not capture the evolution of the species. This effect could also contribute to the different topologies found in previously published phylogenies of the AP complex (Figure 1.10) (Foley et al. 1998, Beebe et al. 1999a, Beebe et al. 2000a). It is also possible that by only including a portion of the AP sibling species in this analysis the establishment of well-supported species relationships in the phylogenetic trees may be hindered. Perhaps by obtaining multiple samples of other remaining AP species help to solidify well-supported relationships within the phylogenetic tree. These additional species are not as widespread and common as AP, AK, AFs.s., AH and AF4. Additionally some of these species are not commonly found in PNG but rather in neighboring Solomon Islands (An. farauti 7/An. irenicus) and Northern Australia (An. farauti 3), both outside of the study locations for our laboratory collaboration. Therefore obtaining samples for this dissertation unfortunately was not possible.
3. Observations of ‘rogue’ individuals

Additional observation within NJ and ML tree results (an later in RAxML results) revealed rogue sequences—individuals found in clades different from their identified species—in both mtDNA trees. Nine rogue individuals were identified in the COI tree (both NJ and ML analysis). Four of these individuals were shown to associate with their own species in Cytb analysis, four individuals were not included for Cytb analysis and one sample (AF2_nal_01, an individual previously identified as An. hinesorum [AH]) was found to associate with the AP clade in both COI and Cytb trees. These rogue observations were consistent with results in Chapter 5 where close examination of mtDNA sequences identified the same disparate individuals as ‘outliers’ in the mismatch distribution calculations of each species.

Four of the 9 rogue individuals were identified in the Cytb tree (both NJ and ML), including AF2_nal_01; DNA from the 5 remaining mosquitoes was no longer available. Two of these individuals, identified as AP were found outside the AP clade (in the AK or AL clades) in the Cytb tree, but were found to associate with the AP clade in COI analysis. AF2_wip_01, an An. hinesorum (AH) individual from Kuru, was a rogue individual that did not associate with any clade, but had previously been found in the AH clade in the COI tree (with low support 60 bootstrap) and VGSC tree (high support 94 bootstrap). Because only 9 disparate sequences, specifically those found to be disparate and associating with the same clade of both mtDNA trees, are observed it is unclear how reliable these observations are. As described in Chapter 5, these sequences were subjected to duplicate, and sometimes, triplicate sequencing before further inclusion in genetic analysis alignments. Therefore
without further observations it cannot be determined if these observations are due to mitochondrial introgression, belong to other members within the species group which \( COI \) has not be reported or if these observations are due to field collection practices where members of the same morphological identification are sometimes stored in the same eppendorf tube prior to DNA extraction. Furthermore, all of these rogue individuals had previously been sequenced for at least one of the nuclear genes and/or identified by molecular species diagnostic tools (described in Chapters 3 and 4: ITS2 LDR-FMA, \( VGSC \) LDR-FMA). In each of these other analyses their species identification was consistent with the species identifier and not consistent with their rogue behavior.

4. **Phenotypes and phylogenetic relationships**

Previous saline tolerance evaluations of some species within the AP complex have identified AFs.s. as the only saline tolerant species (Sweeney 1987), a trait that has allowed this species to exploit the coastal regions throughout PNG mainland, PNG islands (Cooper et al. 2002), the Solomon Islands (Foley and Bryan 2000) and Vanuatu (Reiff et al. 2007). AFs.s. and \( An. \) farauti 7 (AF7) were found to be closely related in NJ analysis (Figure 6.4) with high support (95 bootstrap). This observation was also present in previous phylogenetic reconstructions (Figure 1.10) (Foley et al. 1998, Beebe et al. 1999b, Beebe et al. 2000d). Additionally, in all these reconstructions AH is found to also be closely related to the AFs.s./AF7 clade. Foley and Bryan use this phylogenetic relationship to extrapolate origins of saline tolerance. They report that AFs.s., AH and AF7 are the only members of the Farauti group found in Guadalcanal, Solomon Islands. Larvae of AH and AF7 have only been
recorded from freshwater, whereas AFs.s. larvae are found along the coast in saline pools. Because AFs.s. and AF7 consistently are found to associate with one another phylogenetically, they tested saline tolerance of AF7 and found that, even though larvae had not been found in saline environments, AF7 still had the ability to tolerate saline levels similar to AFs.s. This same evaluation of AH in Guadalcanal determined that it was not tolerant to saline, therefore suggesting that perhaps AF7 had evolved from AFs.s. after it acquired the saline tolerance trait (Foley and Bryan 2000). This is an example of how phylogenetic information was used to predict or test for shared traits among individuals in a tree.

Data generated in this chapter may be used to map other characteristics of these species onto the phylogeny in the future. Vectorial capacity of individual AP sibling species is yet not understood. However feeding trials with *Plasmodium*-infected blood are beginning to come to fruition. Additionally fine-scale biting habits of these individual species are being assessed in conjunction with surveys following LLIN distribution. Having a robust phylogeny of these species may help predict certain characteristics and behaviors of more closely related species within the complex—a useful tool for eliminating disease transmission in endemic communities.

5. **Divergence time of AP sibling species**

Bayesian phylogenetic analysis was used to construct a chronogram estimating divergence times between the species. Divergence between *Anopheles* and *Drosophila* has been estimated at ~260 MYA using molecular clock methods calibrated using fossils from other insect orders (divergence of cockroaches and
crickets/locusts) (Gaunt and Miles 2002). Additionally, most recent common ancestors of *Anopheles* and *Aedes/Culex* have been dated to between 145-200 MYA (Krzywinski et al. 2006, Moreno et al. 2010). By utilizing the *COI* of these species representatives and previously published divergence dates as calibration points, further divergence estimates within *Anopheles* (including AP sibling species) was extrapolated using Divergence-Time Estimation with BEAST.

BEAST analysis suggests that AH diverged from the most recent common ancestor of the AP complex ~6.7 MYA, followed by AK (~3.8 MYA) and AP (~2.23 MYA). This analysis has also suggested that AFs.s. and AF4 diverged most recently (~1.5 MYA). In this analysis the date of estimated divergence for *An. gambiae* and *An. funestus* is approximately 93 MYA and previously published estimates predict this divergence to be around 70-80 MYA (Krzywinski et al. 2006). This analysis was performed using one mitochondrial gene (*COI*), whereas previous dating estimates of *Anopheles* species has used entire mitochondrial genomes. Dating estimates and node confidence within human lineages have been improved by obtaining full-length mitochondrial genomes (Ingman et al. 2000, Ingman and Gyllensten 2003). Therefore by sequencing the entire mitochondrial genomes of AP complex members and including more than the five species described here, we may be able to achieve an improved accuracy of dating estimates and support for nodes.

Interestingly, these dating estimates, especially the approximate time of AH divergence (~6.7 MYA) corresponds to geological events occurring over the last 5-15 million years. Continental plate movement in the area, including collisions between the Sahul plate (Australo-Papuan landmass) and the Sunda plate
(Southeast Asian landmass), occurred in the last 15 million years (Veevers 1991, Johnson 2009). This collision coupled with smaller southward-moving tectonic plates, encroaching upon the northern coast of the New Guinea island in the last 5 million years, have contributed to the formation of the distinct range in altitudes present on the island today, principally the Central Ranges (Hamilton 1979, Veevers 1991, Polhemus and Polhemus 1998, Johnson 2009).

In summary, these phylogenetic data provide a snapshot of the population history of five species within the AP species complex as well as potential historical associations with other extant Anopheles species groups. The results from phylogenetic data presented in this chapter coupled with genetic analysis of the four genes in previous chapters (Chapters 3, 4 and 5) support hypotheses that the mosquitoes analyzed in this dissertation can be categorized into five distinct clades. These observations support the multiple species hypothesis outlined in the aims of this dissertation. The phylogeographical patterns detected in this chapter yield new insight into An. hinesorum populations within mainland PNG. Phylogeographical observations have not been reported in the literature for this species. Therefore further sampling of An. hinesorum within mainland PNG and surrounding islands, where is An. hinesorum known to exist, is warranted to expand upon these observations and improve our understandings of this widespread species.
VII. CHAPTER 7

Project Summary
A. **Summary of Project Findings**

*Anopheles* mosquitoes are vectors of debilitating and deadly disease-causing pathogens and most notorious for being the sole vectors of the *Plasmodium* parasites, causing 225 million cases of human malaria and approximately 800,000 deaths per year (WHO 2010). Almost one third of confirmed, reported cases of malaria in the Western Pacific region occur in Papua New Guinea (PNG), with 100% of the island's population living at risk (95% living at high risk) (WHO 2010). The debilitating disease lymphatic filariasis is also very common in regions of PNG. The filarial parasite causing this disease is also thought to be primarily transmitted by the same malaria-vector *Anopheles* species in PNG (Bockarie et al. 1996a, Bockarie et al. 1998, Bockarie et al. 2002, Cooper et al. 2009a). Malaria eradication campaigns, sometimes thought of as *mosquito* eradication campaigns, of the early and mid 20th century failed to permanently eliminate the vector and interrupt transmission of *Plasmodium* in most tropical and subtropical regions (WHO 1969, Spielman and D’Antonio 2001, Mendis et al. 2009). Given these facts it becomes necessary to better understand the *Anopheles* vectors responsible for disease transmission, especially given the recent renewal of interest in worldwide malaria eradication (WHO 2006b, Gates and Gates 2007).

In Papua New Guinea members of the *Anopheles punctulatus* (AP) species complex are the most widespread and common *Anopheles* species and are thought to be the main vectors of malaria (Cooper et al. 2009a). Of the 12 known AP species complex members, 9 are found within PNG (Cooper et al. 2002). This dissertation has focused on five AP group sibling species. Four species, *An. punctulatus* (AP), *An.
*koliensis* (AK), *An. farauti* s.s. (AFs.s.) and *An. hinesorum* (AH), are common and generally widespread throughout most of the country (Figure 1.9). *An. farauti* 4’s (AF4) range is restricted to specific locations in PNG (Figure 1.9). PNGIMR/CWRU study locations occur in these regions where AF is commonly collected (Benet et al. 2004b). Each of these species exhibits unique but overlapping habitat preferences, therefore it is not uncommon to collect multiple species from the same breeding location or baited trap (Cooper et al. 2002).

Morphological species identification (Figures 1.6 and 2.3) cannot delineate members of the Farauti group (*An. farauti* s.s., *An. hinesorum*, *An. farauti* 3-8) and misclassification of AP and AK as one another or as a member of the Farauti group is common, especially in overlapping habitat areas (Cooper et al. 2002, Benet et al. 2004b, Henry-Halldin et al. 2011). Allozyme analysis and limited cross-mating studies have suggested the different AP complex members were distinct biological species (Bryan 1973a, Bryan 1973b, Mahon and Miethke 1982, Foley et al. 1993). Given that these species demonstrate different habitat preferences (described in Chapter 1, Figure 1.9) it is likely they also differ in other defining characteristics such as host preference, indoor versus outdoor biting and resting habits, vector competency, mating habits, and environmental tolerances. All of these characteristics have potential to influence the mosquito’s exposure to human pathogens and potential to encounter disease control interventions. Therefore having a better understanding of the AP species complex genetic structure improves our understanding of species relationships and reliability of the current species definitions. Ultimately, more positive species identification improves the ability to
determine each species’ importance as a vector of disease. The purpose of this dissertation has been to gain a better understanding of the AP complex by defining genetic diversity within and among sibling species. Results and implications of this effort will be described here.

1. Genetic diversity supporting multiple species hypothesis and defining a baseline understanding of AP species diversity and differentiation

Chapters 3 and 4 of this dissertation highlighted analysis of two nuclear sequences. The ITS2 is commonly used for cryptic species identification in many organisms (Yao et al. 2010). The VGSC encodes for the voltage gated sodium channel, a known target of DDT and pyrethroid insecticides. Both DNA sequences supported the assumption that these five species are distinctly different at the genetic level. Sequence alignments demonstrated species-specific polymorphisms (Figures 3.3 and 4.2) that were then exploited for post-PCR species identification tool development (LDR-FMA).

Complementary mtDNA analysis in Chapter 5 revealed similar support for the multiple species hypothesis suggesting that these five species were distinctly different at mitochondrial loci as well as nuclear loci. However observations, unique to mtDNA analysis, found a few individuals that were previously identified by molecular methods (ITS2 DNA sequence and/or LDR-FMA methods) as one species, to associate with a species different from their identification (Figures 5.1, 5.2, 5.3). These sequences were labeled as ‘outliers’ and were directly responsible for the multiple peaks observed in the mismatch distributions for each species (Figure 5.4)
outliers were also observed to associate with different species clades in phylogenetic analysis (Figures 6.1, 6.5, and 6.6A). It is unclear from the analysis completed here if these observations are artifacts from sample storage methods or possible evidence of mitochondrial introgression or incomplete lineage sorting. Further analysis on larger sample sets targeting mitochondrial genes such as COI or Cytb would be needed to resolve this question. Because mitochondrial targets are easier to amplify than most nuclear genes (multiple copies exist within a cell), mitochondrial genes may be targeted for species identification. However these observations suggest that COI may not be a sufficient gene marker to reliably identify species.

Because 111 individual mosquitoes were sequenced for at least one of the four genes included in this dissertation, observations of intra-species diversity have been possible. Previously, intra-species genetic diversity for the AP group species been reported twice—once for An. farauti s.s. collected from various coastal locations of northern Australia, PNG mainland, PNG island provinces, as well as the Solomon Islands and Vanuatu (Beebe et al. 2000c) and a second study involving An. farauti s.s. and An. irenicus (formerly An. farauti 7) collections on Guadalcanal (Hasan et al. 2008). Geographically unique differences were observed for both studies. Beyond these studies of AFs.s. and An. irenicus, intra-species diversity had not been discussed other than to mention the presence or absence of slightly different RFLP banding patterns among species (Beebe and Saul 1995, Beebe et al. 1999a, Benet et al. 2004b).
This dissertation described intra-species mitochondrial gene diversity of five species in the form of pairwise percent identity as well as heterozygosity, segregating sites and mismatch distributions in Chapter 5. A majority of the individual mosquitoes represented a unique mitochondrial haplotype and heterozygosity ($H$) values were all approaching or equal to 1, suggesting a high level of gene diversity within each species. These results echoed observations within nuclear sequence alignments (Tables 3.2 and 4.5) especially within the coding region of $VGSC$ where a variety of fixed and non-fixed, synonymous mutations were observed within individual species (Table 4.4).

Intra-species phylogeographic patterns were observed in Chapter 6 within the mtDNA genes of AFs.s. and AH clades (Figures 6.1, 6.5 and 6.6). *An. farauti* s.s. members were collected from locations along the northern coast of PNG as well as Manus Island Province, 340 km from the northern coast of PNG. In Neighbor Joining (NJ) and Maximum Likelihood (ML) analysis of COI, the Manus samples (AF1_lor_01, 06, 08, 10) formed a sub-clade within the AFs.s. species clade with bootstrap support >97. Phylogeographical division was observed between AH collected north of the Central Ranges and AH collected in Kuru, south of the Central Ranges. This division was observable in mtDNA phylogenetic trees (NJ, ML and also RAxML), where two clades were supported by BS >84 suggesting that the mountain range is presenting a physical barrier for gene flow between these populations of AH. However, only one location from south of the Central Ranges was sampled. To validate if this phylogeographic division is observed generally between populations on either side of the physical divide may require AH sampling in other locations,
especially south of the mountain range. AH has been reported to exist on islands surrounding PNG, the Solomon Islands and northern Australia (Foley and Bryan 2000a, Beebe and Cooper 2002, Cooper and Frances 2002). It would be interesting to incorporate AH from these locations to further investigate phylogeography of this species.

Generally, phylogenetic analysis revealed distinct, well-supported clades for each species. This provides continuing support of the multiple species hypothesis with respect to the five species included for analysis. Phylogenetic gene trees however did not have a single or well-supported consensus indicating ancestral relationships within the AP species analyzed. AFs.s. and AH were found to be sister taxa with high support in all of the trees, and for some genes AK was found to be closely related to AFs.s. and AH. Further ancestral relationships involving AF4 and AP are not clear.

Divergence time estimates were calculated within the AP complex using previously published molecular divergence estimates within the Diptera Order (Gaunt and Miles 2002) and within Culicidae Family (Krzywinski et al. 2006, Moreno et al. 2010). Results of this analysis suggest that member of the AP complex diverged over the last 7 million years corresponding to the tectonic collisions that have created most of the terrain on the mainland of PNG (Central Ranges and smaller coastal mountain ranges). Therefore it is likely these Anopheles species have inhabited PNG long before humans colonized the island. Nevertheless, Cooper and colleagues suggest these species are generally unspecialized in their habits and have adapted to living in cohabitation with humans. This observation can be seen
most drastically with *An. punctulatus*. With a growing number of motor vehicles in the country this species has found wheel tracks, where fresh rainwater collects on a temporary basis, along with drains surrounding human establishments to be the most suitable larval habitat (Cooper et al. 2002).

2. **Molecular tools**

Chapters 3 and 4 also discussed development and evaluation of two post-PCR high-throughput molecular diagnostic assays for species identification, one focusing on a traditional cryptic species DNA target, ITS2 and the other on polymorphic intron regions of the *VGSC*. Field collections representing 7 PNG provinces were successfully evaluated using these LDR-FMA methods demonstrating their capability to assign species identification in an unbiased, high-throughput manner. Both of these assays were highly concordant (>98% concordant) with the molecular gold standard for AP species identification, ITS2 PCR-RFLP. Apart from technical advantages discussed in Chapters 3 and 4, the LDR-FMA strategies developed here introduce many advantages of multiplex analysis.

The practical utility of ITS2 LDR-FMA was tested in an evaluation of a field collection (n=868 mosquitoes) conducted during a parasite-monitoring survey of an area in East Sepik Province where lymphatic filariasis is transmitted in high densities (Bockarie et al. 1998, Thomsen et al. 2010). A portion of this sample set had been previously subjected to ITS2 PCR-RFLP species identification and LDR-FMA results were >95% concordant with these previous molecular results. This demonstrated the field application of this assay. Throughout work with the ITS2 PCR and LDR-FMA, it was also possible to skip the DNA extraction step by adding a
piece of the mosquito’s leg directly to the ITS2 PCR. This method did not interfere with LDR-FMA species diagnosis and would be useful in resource-limited situations where DNA extraction is both time consuming and costly.

The VGSC LDR-FMA proved to be a uniquely informative tool as it has the capability of identifying species while simultaneously determining if mutations associated with insecticide resistance (kdr) are present (heterozygous mutation carrier or homozygous resistant) or absent (homozygous sensitive). This assay's ability to identify species was compared with the ITS2 LDR-FMA and found to be >99% concordant. Additionally all individuals assayed in this dissertation lacked the resistance-associated mutations. This assay will become useful in future field evaluations of the country-wide distribution of LLINs (Hetzel 2009) where mosquitoes may be under intensive insecticide pressure. These results suggest that the mosquitoes in pre-LLIN populations do not carry kdr-associated mutations. Monitoring the species composition as well as for the development resistance mutations in the population will be necessary to ensure continued success of the current vector management strategy.

Validation of these two LDR-FMA assays present general results that can be interpreted two different perspectives—these tools (1) demonstrated a useful utility for species surveys in their ability to efficiently differentiate the five AP species but they also demonstrated (2) the durability of the multiple species hypothesis. In this respect the reliability of the genetic species definitions was evaluated by probing for polymorphisms that were observed to occur in a species-specific manner (Figures 3.4 and 4.3). This perspective can be used to continue
evaluating mosquito populations and whenever the LDR-FMAs are unable to identify an individual, the sample should be sequenced as it may harbor an uncharacterized polymorphism or represent a species that has not been included in the assay. This leads to another advantage of the LDR-FMA approach: additional DNA probes can be added to the assay very easily and at a low cost. Therefore, the assays can be expanded to probe for other *Anopheles* species or even specific SNPs of interest, for instance those that may delineate strains within a mosquito species.

B. Integrated Vector Management: Contributions of this Project

Integrated Vector Management (IVM), as described by the WHO, is a “rational decision-making process for the optimal use of resources in the management of vector populations, so as to reduce or interrupt transmission of vector-borne diseases” (WHO 2004). Many factors contribute to an IVM plan, however the most influential is the local ecology of the disease vector. A high-resolution understanding of the vector's ecology may be needed to identify vector control measures appropriate to the specific vector population and design a public health plan that will be successful in interrupting disease transmission. For example, if the local vector is characteristically an early evening, outdoor biter, LLINs may have little to no effect on reducing disease transmission. Or if the local vector has a tendency to bite indoors but then leave the dwelling to rest outside the home, IRS may not be the most appropriate or successful method to interrupt disease transmission.
Very recently malERA (malaria Eradication Research Agenda) Consultative Group, sponsored by Roll Back Malaria (RBM) (a partnership between the WHO, United Nations Development Program, UNICEF and the World Bank), has outlined the vector control interventions needed for sustained control of malaria and for malaria eradication (malERA 2011). Appropriate IVM is clearly needed for sustained control and eradication of malaria, and these principles can be translated to elimination of other vector borne disease. In planning for control or eradication goals, malERA recommends that new vector control strategies explored through modeling and if theoretically successful move to field trials in the endemic setting.

1. **IVM in PNG**

In order to plan for such vector control exploration in PNG, whether theoretical or in the field, a firm understanding of the AP species complex ecology is needed—transcending the known information regarding larval habitats and mosquito habitat range. Fine scale biting (anthropophily, biting times, indoor vs. outdoor biters etc.), resting, oviposition and mating location preferences will be necessary to identify the most appropriate vector control measures for specific areas. Additionally it will become necessary to understand the ability each species in the AP species complex has to acquire, sustain and propagate a *Plasmodium* infection (vector competency), as it is known that not all *Anopheles* species are capable of transmitting malaria (Service 2002). Gathering this information about the AP complex requires first and foremost, clear species definitions and secondly, a method of accurate and reliable species identification.
This dissertation has sought to provide a more extensive understanding of the molecular species definitions of the most common AP complex species within PNG in addition to describing reliable high-throughput species identification methods designed to evaluate large field collections. Since data suggests that the five species analyzed are distinctly (genetically) different we can assume that they are independent species capable exhibiting completely different ecologies and therefore may contribute differently to disease transmission within a population. It would also be important to understand if the habits of a species were consistent across its entire range, or if they were subject to local variation. This would be crucial to the sustained success of vector management at the local level.

Sequence comparisons and phylogenetic analysis suggest that members of the AP complex are continuing to diversify. Phylogeographical patterns suggest that geographically separated populations within An. farauti s.s. and An. hinesorum are genetically different. Haplotype data for An. koliensis suggest that diversity is high within this species. None of the 19 AK sequences shared the same COI haplotype and a majority of these collections were made in locations that were separated by less than a kilometer. It is unclear if this diversity affects the organisms vectorial behavior and how much time it has taken for this diversity to accumulate in these populations. Understanding how quickly mosquitoes are able to adapt/evolve will become important in designing vector management approaches that can be successfully completed before the vector has time to adapt.

Sequence and phylogenetic comparisons in this dissertation have also suggested that An. punctulatus, An. koliensis, An. farauti s.s., An. hinesorum and An.
*farauti* 4 are independent species in which heritable traits will be acquired and transmitted within each species independently. These observations may extend to the geographically isolated populations discussed above as well. Given LLINs have been distributed to the entire country (Hetzel 2009) it is necessary to monitor for insecticide resistance in every species within the AP complex as well as within geographically distinct populations of the same species. This is not welcome news for vector management groups who may be looking to cut back on expensive monitoring.

Ecological preference has been reported to differ between strains of other *Anopheles* species. For example, oviposition preference of *An. gambiae* was shown to differ between the two chromosomal forms of the species, where S-form prefer temporary fresh rainwater pools lacking predators and M-form prefer less temporary, predator-rich pools (Lehmann and Diabate 2008). These molecular forms are known to hybridize in the wild (Lawniczak et al. 2010), however it is unclear what oviposition sites hybrids prefer due to deficient larval sampling of hybrids (Besansky 2010 personal communication). The data presented in Chapters 5 and 6 suggest that mitochondrial introgression may be occurring between members of the AP complex. If this observation is validated through large-scale surveys it will be important to understand how hybrid populations behave with respect to their ecological preferences and vector competency. Additionally, at this point it is unclear if strains of AP complex species exist, underscoring the need for continued investigation of the species complex. The LDR-FMA methods described in
this dissertation provide a means to screen large mosquitoes surveys possibly allowing for identification of genetically distinct strains within individual species.

2. **IVM for other *Anopheles* species complexes**

   To date, greater than 30 *Anopheles* species complexes have been described, most of which include more than 4 identified sibling species. These species complexes contribute to over 25% of the currently identified *Anopheles* species (Harbach 2004, WHO/SEARO 2007). A majority of these *Anopheles* complexes contain species that are thought to be dominant vectors, potentially important (Kiszewski et al. 2004) or at least capable of transmitting malaria (Baimai 1988, Amerasinghe et al. 1992, Carnevale et al. 1992, Kitron et al. 1994, Bryan et al. 1996, Proft et al. 1999, Conn et al. 2002, Prakash et al. 2005). Since so many sibling species groups are implicated as vectors of malaria and most of these groups are found in areas where malaria is still endemic, establishing successful IVM plans will require extensive fine-scale knowledge of the vector population and sibling species ecological differences.

   Interestingly, the AP species complex is the largest species complex (in terms of number of species) described thus far (Harbach 2004, WHO/SEARO 2007) and therefore perhaps the most challenging to study. This dissertation provides the organizational framework for deciphering differences among AP complex members. However, the same practices outlined here could also be applied to other understudied *Anopheles* species complexes in order to establish baseline species information leading to more accurate species identification, ecological descriptions and thus improved IVM strategies.
C. **Malaria Elimination and Importance of Molecular Epidemiology**

In late 2007 The Bill and Melinda Gates Foundation called for malaria eradication (Gates and Gates 2007)—*complete* and permanent elimination of the *Plasmodium* parasite worldwide (WHO 2006b, Mendis et al. 2009). Since then, collaborative initiatives like malERA have identified areas where further research is needed to achieve sustained control and ultimately eradication of the parasite. One major requirement for both successful disease control and eradication is accurate species identification and incrimination as a vector of disease.

As the work towards malaria control and elimination progresses, molecular epidemiology tools designed to detect *Plasmodium* parasites both in the human and *Anopheles* vector populations will become increasingly more important. Studies from the late 1980s-early 90s in areas around Madang (Madang Province), where approximately 40% of the surveyed human population was infected with *Plasmodium* (Cattani et al. 1986), demonstrated that only 1-3% of captured mosquitoes were positive for *Plasmodium* via a circumsporozoite antigen detection method (CSP ELISA) (Graves et al. 1990). A similar observation was made in Western Province where 1-2% of captured mosquitoes were positive using CSP ELISA, however this area was reported to have a lower endemicity than Madang (Hii et al. 1997). This infection rate in the mosquito is a common observation for vectors in other malaria endemic regions of the world and can vary from <1-5% between regions, seasons, and vector species (Beier 1998, Service 2004). Given this intrinsically low infection rate in the mosquito population, employing sensitive and
high-throughput techniques to monitor for *Plasmodium* in the vector populations will become necessary as control measures are implemented.

A future direction of the work presented in this dissertation would be to develop an extension of the LDR-FMA *Anopheles* species identification method to include detection of *Plasmodium* parasites. As shown for *Wuchereria bancrofti*, this approach, when used to evaluate a field collection of mosquitoes, has demonstrated the ability to simultaneously differentiate AP species while determining *W. bancrofti* infection status of the individual mosquitoes. This analysis revealed ~20% of the surveyed mosquito population was positive for *W. bancrofti* which is consistent with previous laboratory observations in this survey location where mosquitoes were evaluated using agarose gel band detection of PCR products (Thomsen et al. 2010). This exercise demonstrates the ITS2 LDR-FMA’s capacity to be expanded for concurrent detection of pathogens harbored by *Anopheles*, useful in epidemiology field surveys and xenomonitoring.

A second major requirement for both successful disease control and eradication is insecticide susceptibility and monitoring. malERA calls for improved vector monitoring to target appropriate interventions for vector management (malERA 2011). Monitoring phenotypic responses to insecticides is necessary to detect resistance in the vector population. This type of monitoring requires rearing wild-caught larva into adults, subjecting females to insecticide treatments, monitoring them for a period of time and determining their response phenotype (WHO 2006a). Additionally, insecticide bioassays are not designed to assign resistance phenotypes of mosquitoes that are captured in the field, since there is no
way of delineating death from insecticide exposure or other causes, either natural or
associated with capture. Bioassays therefore are laborious, do not have the efficient
capacity of analyzing large collections, and may not detect early stages of resistance
development associated with genetic mutations in a population. Therefore
molecular methods, capable of surveying large populations for common mutations
associated with insecticide resistance, will be useful to detect populations where
resistance may be developing and inform vector management strategies (Ranson et
al. 2011). The VGSC LDR-FMA will serve as a useful tool for this purpose especially
in follow-up surveys associated with countrywide distribution of LLINs. The design
of this evaluation tool can also be modified to target mutations associated with
resistance in other malaria endemic regions of the world.

D. Future Directions of Research Associated with this Project

Through collaborative effort between our laboratory and a group at the
Cleveland Clinic Foundation, the complete genomes of *An. punctulatus* and *An.
farauti* s.s., *An. hinesorum* and *An. koliensis* have just recently been sequenced.
Interestingly, in preliminary assembly of the genomic data, it appears that the AP
complex species share only ~2% homology (unpublished data) with the *An.
gambiae* genome that was sequenced in 2002 (Holt et al. 2002). This demonstrates
the extreme difference between species especially those belonging to the same
subgenus (*Cellia*). Because so much information is known about *An. gambiae* (Table
1.1) it is often thought of as the model organism for malaria vectors. This genome
comparison, however, underscores potential for significant differences and the need
to continue research of the AP species complex as generalities derived from research with *An. gambiae* may not translate to the AP complex given their genetic dissimilarity with *An. gambiae*.

Sequence data generated in this dissertation supports the existence of genetically distinct species. Furthering this sequence data to whole genome sequencing for all AP complex members will open up seemingly endless amount of research opportunities within this species complex. Genome sequences will enable analysis to determine more accurate species relationships and ancestry, conclusions that could not be achieved using multiple genes in this dissertation. Additionally, if whole genomes are obtained for many AP complex members it will allow for identification of genes that can be associated with ecological traits. For example, perhaps there is a genetic reason for the shared saline tolerance trait of *An. farauti* s.s. and *An. farauti* 7. If a specific species is identified as being more capable of *Plasmodium* transmission than others, we can pursue identification of genes associated with this trait.

malERA suggests novel approaches, such as genetic modification, need to be explored in order to reduce the vectorial capacity of primary disease vectors (malERA 2011) for *Plasmodium* eradication purposes. Identifying the genetics behind behavioral or ecological traits will assist targeted vector management strategies and may even aid in designing competitive laboratory strains of these species that are, for example, resistant to *Plasmodium* infection or incapable of producing viable offspring. However, given the diversity of species that exist in this
lesser region of the malaria endemic world, it is questionable if genetic modification approaches would be successful.

Genome sequencing will also provide SNP data that can be used to design assays detecting hybridization between AP complex species. Since hybridization has not been definitively ruled out among members of the AP complex it would be important to detect successful hybridization in the wild to assess possible spread rate of advantageous mutations. If a mutation, such as kdr, were found in a mosquito population known to successfully hybridize with another species this would increase the spread rate of this mutation beyond the expected rate within the original species—causing concern for IVM success.

While genome sequencing is a promising tool for the malaria eradication effort it is costly, time consuming, and may not be complete for a couple of years. Therefore SNPs identified among the four studied genes in this dissertation are able to be immediately useful in evaluating the diversity of vectors in a population. Additionally, our lab has recently assembled the *Wuchereria bancrofti* mitochondrial genome and SNPs have been identified in many genes. Since *W. bancrofti* DNA can be amplified out of an individual mosquito, association studies can be conducted between parasite and mosquito genotypes, identifying the vector capacity of these species, or even identification of strains within a species.

In conclusion, it has become clear that members of the AP complex, previously identified as the main vectors of malaria in PNG (Cooper et al. 2009a), are closely related genetically distinct species. Understanding how this genetic differentiation affects biological phenotypes such as vector competence or breeding
habitat preference needs to be addressed. The experiments described in this dissertation do not seek to address these questions, but have provided the initial framework needed to begin genetic-based ecological studies that will further improve Integrated Vector Management strategy with the goal of eradicating human malaria.
VIII. Appendix A

Sequence Alignments of Four Genes Used in this Dissertation
Figure A.1 Internal Transcribed Spacer 2 (ITS2) alignment. Ninety individual AP complex representatives (Table 3.2) and an *An. longirostris* representative as a comparison (sequence #91) were used in this dissertation. Image is a zoom-out of the ClustalW alignment in Geneious, where black bars represent regions of polymorphism.
Figure A.2 Alignment of the portion of the Voltage Gated Sodium Channel (VGSC) used in this dissertation. Ninety individual AP complex representatives (Table 4.2) and an *An. longirostris* representative as a comparison (sequence #91) were used in this dissertation. Image is a zoom-out of the ClustalW alignment in Geneious, where black bars represent regions of polymorphism.
Figure A.3. Cytochrome Oxidase I (COI) alignment. Ninety-seven individual AP complex representatives (Table 5.3 Total) and an *An. longirostris* representative as a comparison (sequence #98) were used in this dissertation. Image is a zoom-out of the ClustalW alignment in Geneious, where black bars represent regions of polymorphism.
Figure A.4 Cytochrome b (Cytb) alignment. Thirty-seven individual AP complex representatives (Table 5.4) and two *An. longirostris* representative as a comparison (sequence #38 and 39) were used in this dissertation. Image is a zoom-out of the ClustalW alignment in Geneious, where black bars represent regions of polymorphism.
IX. References


Beck, J., I. Kitching, and K. Linsenmair. 2006. Wallace’s line revisited: has vicariance or dispersal shaped the distribution of Malesian hawkmoths (Lepidoptera: Sphingidae)? Biological Journal of the Linnean Society 89:455-468.


Beebe, N., R. Cooper, D. Morrison, and J. Ellis. 2000a. A phylogenetic study of the *Anopheles punctulatus* group of malaria vectors comparing rDNA sequence
alignments derived from the mitochondrial and nuclear small ribosomal subunits. Molecular Phylogenetics and Evolution 17:430-436.

Beebe, N., J. Ellis, R. Cooper, and A. Saul. 1999a. DNA sequence analysis of the ribosomal DNA ITS2 region for the Anopheles punctulatus group of mosquitoes. Insect Molecular Biology 8:381-390.

Beebe, N. W., B. Bakote'e, J. T. Ellis, and R. D. Cooper. 2000b. Differential ecology of Anopheles punctulatus and three members of the Anopheles farauti complex of mosquitoes on Guadalcanal, Solomon Islands, identified by PCR-RFLP analysis. Medical and Veterinary Entomology 14:308-312.


Beebe, N. W., J. T. Ellis, R. D. Cooper, and A. Saul. 1999b. DNA sequence analysis of the ribosomal DNA ITS2 region for the Anopheles punctulatus group of mosquitoes. Insect Molecular Biology 8:381-390.


Davies, T. G., L. M. Field, P. N. Usherwood, and M. S. Williamson. 2007. DDT, pyrethrins, pyrethroids and insect sodium channels. IUBMB Life **59**:151-162.
Djouaka, R., A. Bakare, O. Coulibaly, M. Akogbeto, H. Ranson, J. Hemingway, and C. Strode. 2008. Expression of the cytochrome P 450 s, CYP 6 P 3 and CYP 6 M 2 are significantly elevated in multiple pyrethroid resistant populations of Anopheles gambiae s.s. from Southern Benin and Nigeria. BMC Genomics 9:538.


Miller, M. H., MT; Vos, R; Midford, PE; Liebowitz, T; Chan, L; Hoover, P; Warnow, T. 2009. The CIPRES Portals V 3.1. CIPRES (Cyberinfrastructure for Phylogenetic Research). http://www.phylo.org/sub_sections/portal


Rambaut, A. 2009. FigTree v1.3.1, a graphical viewer of phylogenetic trees. See http://tree.bio.ed.ac.uk/software/figtree/, Edinburgh.


WHO/SEARO. 2007. Anopheline Species Complexes in South and South-East Asia, SEARO Technical Publication No. 57. WHO Regional Offices for South-East Asia, New Delhi.


